Metti⁸

Advanced Autumn School - **8th** Edition -

Thermal Measurements & Inverse Techniques

Sept. 24th – Sept. 29th, 2023 - *Ile d'Oléron, France*

VOLUME 1 : LECTURES

Metti⁸

Advanced Autumn School

Thermal Measurements

&

Inverse Techniques

- 8th Edition -

Sept. 24th – Sept. 29th, 2023
Ile d'Oléron – France

https://metti8.sciencesconf.org/

Supported by :

https://metti8.sciencesconf.org/

Scientific coordination of METTI 8:

Denis Maillet
LEMTA, Nancy
Denis.Maillet@univ-lorraine.fr

Jean-Luc Battaglia
I2M, Bordeaux
jean-luc.battaglia@u-bordeaux.fr

Organisation of METTI 8

Yassine Rouizi
LMEE, Evry
Tel.: (33) 1 69 47 79 31
yassine.rouizi@univ-evry.fr

Olivier Quéméner
LMEE, Evry
Tel.: (33) 1 69 47 79 38
o.quemener@iut.univ-evry.fr

(Secretary) Olivia Viardot
LMEE, Evry
Tel.: (33) 1 69 47 75 51
olivia.viardot@univ-evry.fr

METTI 8 Commitee

J. C. Batsale, J. L. Battaglia, J. G. Bauzin, J. Berger, S. Demeyer, T. Duvaut, Y. Favennec, J.-L. Gardarein, B. Garnier, J. Gaspar, N. Horny, L. Ibos, J. C. Krapez, F. Lanzetta, N. Laraqi, P. Le Masson, C. Le Niliot, D. Maillet, J. Meulemans, H. Orlande, L. Pérez, T. Pierre, O. Quéméner, B. Rémy, F. Rigollet, C. Rodiet, S. Rouchier, Y. Rouizi, P. Salagnac

METTI 8 Location

CNRS Center "La Vieille Perrotine village", Oléron Island, (France)

https://metti8.sciencesconf.org/

FOREWORD

The 8th edition of the Advanced Autumn school 'Thermal Measurement and Inverse Techniques' is run by the METTI Group (**ME**sures en **T**hermique et **T**echniques **I**nverses) that constitutes a division of the Société Française de Thermique (SFT, French Heat Transfer Society).

* * *

Finding 'causes' from measured 'consequences' using a mathematical model linking the two is an inverse problem. This is met in different areas of physical sciences, especially in Heat Transfer. Techniques for solving inverse problems as well as their applications may seem quite obscure for newcomers to the field. Experimentalists desiring to go beyond traditional data processing techniques for estimating the parameters of a model with the maximum accuracy feel often ill prepared in front of inverse techniques. In order to avoid biases at different levels of this kind of involved task, it seems compulsory that specialists of measurement inversion techniques, modelling techniques and experimental techniques share a wide common culture and language. These exchanges are necessary to take into account the difficulties associated to all these fields. It is in this state of mind that this school is proposed. The METTI Group (Thermal Measurements and Inverse Techniques), which is a division of the French Heat Transfer Society (SFT), has already run or co- organized seven similar schools, in the Alps (Aussois, 1995 and 2005), in the Pyrenees (Bolquère-Odeillo, 1999), in Brasil (Rio de Janeiro, 2009), in Bretagne (Roscoff, 2011), in Pays Basque (Biarritz, 2015) and in Porquerolles island (Porquerolles 2019). For this eighth edition the school is again open to participants from the European Community with the support of the Eurotherm Committee.

* * *

Two books are distributed at the beginning of the school. Volume 1 contains the texts used as supports for the lectures and Volume 2 contains the texts used as supports for the tutorials.

https://metti8.sciencesconf.org/

LECTURES - TABLE OF CONTENTS and schedule

Monday, September 25th		Tuesday, September 26th		Wednesday, September 27th		Thursday, September 28th		Friday, September 29th	
8:00 - 8:15	Welcome								
8:15 - 9:00	L1 - Getting started with problematic inversions with three basic examples	8:30 - 10:00	L4 - Measurements without contact in heat transfer	8:30 - 10:00	L6 - Inverse problems and regularized solutions	8:30 - 10:00	L8 - Optimization tools dedicated to function estimation in inverse heat transfer problems.	8:30 - 10:00	Tutorial session 9
9:00 - 10h30	L2 - Advanced measurements with contact in heat transfer: : principles, implementation and pitfalls								
10h30 - 10h50	Coffee Break around posters	10h00 - 10h45	Coffee Break around posters	10h00 - 10h45	Coffee Break around posters	10h00 - 10h45	Coffee Break around posters	10:00 - 10:30	Coffee Break around posters
10h50 - 12:20	L3 - Basics for linear inversion, the white box case	10:45 - 12h15	L5 - Non linear parameter estimation problems: tools for enhancing metrological objectives	10:45 - 12h15	L7 - Types of inverse problems, model reduction, model identification	10:45 - 12h15	L9 - The Use of Techniques within the Bayesian Framework of Statistics for the Solution of Inverse Problems	10:30 - 11:30	Invited Conference : Contactless thermal measurements using MRI, V. Ozenne
								11h45 - 12:30	Closing session - Debriefing
12:20 - 13:30	Lunch	12:15 - 13:30	Lunch	12:15 - 13:30	Lunch	12:15 - 13:30	Lunch	12:30 - 13:30	Lunch
13:30 - 16:40	Free time	13:30 - 16:40	Free time	13:30 - 16:40	Free time	13:30 - 16:40	Free time		
16:40 - 18h10	Tutorial session 1	16:40 - 18h10	Tutorial session 3	16:40 - 18h10	Tutorial session 5	16:40 - 18h10	Tutorial session 7		
18:10 - 18:30	Pause	18:10 - 18:30	Pause	18:10 - 18:30	Pause	18:10 - 18:30	Pause		
18:30 - 20:00	Tutorial session 2	18:30 - 20:00	Tutorial session 4	18:30 - 20:00	Tutorial session 6	18:30 - 20:00	Tutorial session 8		

Lecture 1: Getting started with problematic inversions with three basic examples

P. Le Masson[1], O. Fudym[2], J.-L. Gardarein[3], D. Maillet[4],

[1] Université Bretagne Sud, IRDL UMR 6027 CNRS, Lorient, France
E-mail: philippe.le-masson@univ-ubs.fr
[2] IMT Mines Albi-Carmaux, RAPSODEE UMR 5302 CNRS, Albi, France
E-mail: olivier.fudym@gmail.fr
[3] Aix-Marseille Université, IUSTI UMR 7343 CNRS, Marseille, France
E-mail: jean-laurent.gardarein@univ-amu.fr
[4] Université de Lorraine, LEMTA UMR 7563 CNRS, Vandoeuvre-lès-Nancy, France
E-mail: denis.maillet@univ-lorraine.fr

Abstract. Introduction to the inverse approach is made starting by simple examples (solution of a linear system of equations, with noised right hand member, case of a slab, in steady state regime, with either flux or conductivity estimation). The inverse terminology, the pitfalls of inversion (noise amplification effect), as well as the corresponding methodological approach are highlighted. The objective is not to solve these problems but to pinpoint the main crucial points in inverse measurement problems. Other lectures (L3 & L7 to L10) will be used to show how to solve them, with the help of the points studied in the lectures in between.

Introduction

Inverse problems are part of our daily practice, even if we do not know they are inverse problems. We consider here a scientific field (heat transfer, mechanical or chemical engineering, physics...) where a quantitative model is available, that is a mathematical procedure which is able to simulate, with a good enough precision (the model can sometimes be reduced and therefore offset with respect to the physical problem), the phenomena at stake. The inverse use of this model gives rise to an inverse problem. Instead of introducing the different notions associated to such problems, which will be progressively dealt with in the following lectures of this advanced school, we will present examples that correspond to the inverse use of a model, as well as the specific problems that appear concomitantly. These examples will correspond to *exact matching* between measurements (noted y or Y further on) and model outputs (noted y_{mo} or T or ΔT futher on), with no use of a least square approach. The term "*exact matching"* means that inversion is made through solving an equation where both model outputs and measurements are equal, which is only possible when the number of unknowns is equal to the number of measurements. Consequently, the least square sum is not only minimum but equal to zero.

2. Example 1: square system of linear equations

Let us suppose we have a linear model that allows to get m output values $y_{mo1}, y_{mo2}, \ldots, y_{mom}$ for any values of the input values $x_1, x_2, \ldots, x_m$. Note that we assume here that both numbers

of input and output values are the same and that the output values are subscripted by the index "*mo*" to remind us that it is only a model. It is very convenient to use here column vectors to represent this linear relationship under the form:

$$\boldsymbol{y}_{mo} = \boldsymbol{S}\,\boldsymbol{x} \tag{1.1}$$

where $\boldsymbol{y}_{mo}$ and $\boldsymbol{x}$ are both (m, 1) matrices (column vectors) composed of the y_{mo}'s and of the x's and $\boldsymbol{S}$ a square (m, m) matrix, which is called a « *sensitivity matrix* » in the inverse problem terminology.

In the direct problem input $\boldsymbol{x}$ is known and $\boldsymbol{y}_{mo}$, the output of the model, is calculated.

The example that will be studied here corresponds to the m = 2 case, with:

$$\boldsymbol{S} = \begin{bmatrix} 10 & -21 \\ 39 & -81 \end{bmatrix} \qquad \boldsymbol{x} = \boldsymbol{x}^{exact} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \;\Rightarrow\; \boldsymbol{y}_{mo} = \boldsymbol{S}\,\boldsymbol{x}^{exact} = \begin{bmatrix} 9 \\ 36 \end{bmatrix} \tag{1.2}$$

We have supposed here that, in the given problem, we know the exact value $\boldsymbol{x}^{exact}$ of the input vector $\boldsymbol{x}$.

Conversely, if that is $\boldsymbol{y}_{mo}$ which is known, solution of system (1.2), or inversion of matrix **S**, provides the true value of the input:

$$\boldsymbol{x}^{exact} = \boldsymbol{S}^{-1}\boldsymbol{y}_{mo} \tag{1.3}$$

We have therefore solved the *inverse problem* using exact data $\boldsymbol{x}$.

Let us now assume that the output, that is the data, corresponds to some measurements of $\boldsymbol{y}_{mo}$ which are corrupted by an additive noise $\boldsymbol{\varepsilon} = \begin{bmatrix} 0.1 & -0.3 \end{bmatrix}^T$. Superscript T designates the transpose of a matrix here. Each component of this noise represents about 1%, in relative value, of the corresponding component of the exact output $\boldsymbol{y}_{mo}$:

$$\boldsymbol{y} = \boldsymbol{y}_{mo} + \boldsymbol{\varepsilon} = \begin{bmatrix} 9{,}1 \\ 35{,}7 \end{bmatrix} \tag{1.4}$$

The natural idea for retrieving an approximate solution of the inverse problem is to replace the exact model output $\boldsymbol{y}_{mo}$ by its measured value $\boldsymbol{y}$ in (1.4), or to solve linear system (1.1) $\boldsymbol{S}\,\boldsymbol{x} = \boldsymbol{y}$ with this noised right hand member:

$$\begin{bmatrix} 10 & -21 \\ 39 & -81 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 9.1 \\ 35.7 \end{bmatrix} \tag{1.5}$$

to recover an estimated value of the input

$$\hat{\boldsymbol{x}} = \boldsymbol{x}^{exact} + \boldsymbol{e}_{\boldsymbol{x}} = \begin{bmatrix} 1.40 \\ 0.233 \end{bmatrix} \tag{1.6}$$

$\boldsymbol{e}_x$ designates the error on the $\hat{\boldsymbol{x}}$ estimate.

This means that an error of 53 % has been made for x_1 (1.40 instead of 3) and of 77% for x_2 (0.233 instead of 1). This phenomenon is illustrated in figure 1: two far away values of $\boldsymbol{x}$, $\boldsymbol{x}^{exact}$ the exact value and $\hat{\boldsymbol{x}}$ the solution of (1.4), yield approximately the same values, within $\boldsymbol{\varepsilon}$, in the y_1 - y_2 plane. In this case, the determinant of matrix $\boldsymbol{S}$ is not very close to zero: its value is 9.

Let us note that, in this particular case, this solution $\hat{\boldsymbol{x}}$ of system $\boldsymbol{S}\,\boldsymbol{x} = \boldsymbol{y}$ is also an ordinary least squares solution of model (1.1) with noisy data $\boldsymbol{y}$.

In order to analyse the possibly "pathological" character of the solution of $\boldsymbol{S}\,\boldsymbol{x} = \boldsymbol{y}$, two global criteria, the amplification coefficients of the absolute and relative errors, k_a and k_r, respectively can be introduced. Their values can be calculated, using the Euclidian norm L_2:

$$k_a(\boldsymbol{\varepsilon}) = \frac{\|\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}\|}{\|\boldsymbol{\varepsilon}\|} = \frac{\|\boldsymbol{e}_x\|}{\|\boldsymbol{\varepsilon}\|} = \frac{1.774}{0.316} = 5.61 \quad \text{with} \quad \|\boldsymbol{u}\| = \left(\sum_{j=1}^{2} u_j^2\right)^{1/2} \quad \text{and} \quad \boldsymbol{e}_x = \hat{\boldsymbol{x}} - \boldsymbol{x}^{exact}$$

$$\text{and} \quad k_r(\boldsymbol{\varepsilon}) = \frac{\|\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}\| / \|\boldsymbol{S}^{-1}\boldsymbol{y}_{mo}\|}{\|\boldsymbol{\varepsilon}\| / \|\boldsymbol{y}_{mo}\|} = \frac{\|\boldsymbol{e}_x\| / \|\boldsymbol{x}^{exact}\|}{\|\boldsymbol{\varepsilon}\| / \|\boldsymbol{y}_{mo}\|} = \frac{1.774 / 3.16}{0.316 / 37.11} = 65.8 \tag{1.7}$$

Figure 1 shows the amplification effect of the measurement noise in the above example.

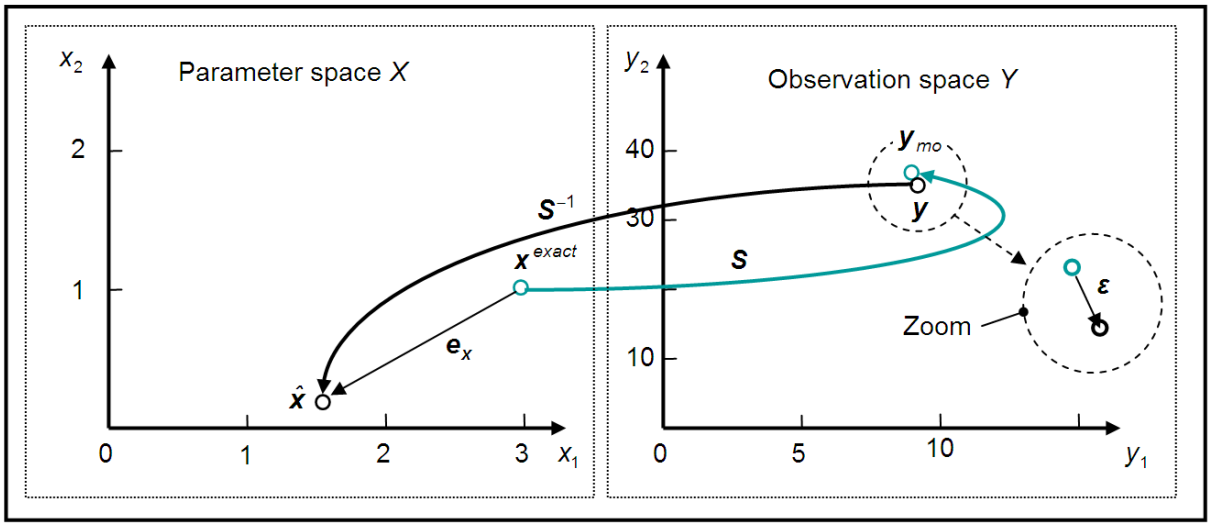


Figure 1 – Effect of the measurement error on parameter estimation through inverse mapping

Criteria (1.7), which measure the amplification effect of the measurement noise $\boldsymbol{\varepsilon}$ allow to quantify the unstable character of the solution. In practice, calculation of these criteria, which requires a prior knowledge of the exact value $\boldsymbol{x}^{exact}$ of the unknown, is not possible. In order to analyze this stability problem, a condition number of matrix $\boldsymbol{S}$ shall be introduced, here for a square matrix.

Remark 1

In figure 1, the exact $\boldsymbol{x}^{exact}$ and estimated $\hat{\boldsymbol{x}}$ values of parameter vector $\boldsymbol{x}$ are shown in the left hand side, in the two-dimension vector space of the *parameters X* (also called *input* space), where an orthonormal basis that corresponds to the components (x_1, x_2) of these vectors has been chosen. In the right hand side, the output $\boldsymbol{y}_{mo}$ of the model, and measurements $\boldsymbol{y}$ are shown in the *observation* space *Y* where a corresponding orthonormal coordinates system (y_1, y_2) has been selected. The two norms present in the definition of k_a are the lengths of the vectors of the estimation error $\boldsymbol{e}_x = \hat{\boldsymbol{x}} - \boldsymbol{x}^{exact}$ and of the measurement noise $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{y}_{mo}$. The other extra norms present in the definition of k_r are the lengths of the vectors representing the exact values $\boldsymbol{x}^{exact}$ (model input) and $\boldsymbol{y}_{mo}$ (model output).

Remark 2

The norms used in (1.7) are not necessarily the same in spaces *X* and *Y*. For example, coordinates (x_1, x_2) can be expressed in W.m^{-2}, if the unknowns are fluxes and coordinates (y_1, y_2) can be temperatures (Kelvin). However, in order to define such norms in each space, x_1 and x_2 should have the same units as well as y_1 and y_2. If it is not the case a scaling has to be implemented in both domains.

Remark 3

Coefficient k_r does not depend on the physical dimensions in *X* and *Y*: it explains the transformation of the noise/signal ratio $\|\boldsymbol{\varepsilon}\| / \|\boldsymbol{y}_{mo}\|$ into a relative estimation error $\|\boldsymbol{e}_x\| / \|\boldsymbol{x}^{exact}\|$. The inverse process, where one starts from the measurement domain *Y* to get a value of the input in the parameters domain *X*, corresponds to the inverse linear mapping $\boldsymbol{S}^{-1}$. Passage from *Y* space into *X* space is associated with a high amplification of the error: this problem is therefore ill-conditioned.

Remark 4

The high value $k_r(\boldsymbol{\varepsilon}) = 65.8$ of the relative amplification coefficient is not the highest possible here: things can become even worse. This maximum value of this coefficient is the condition number (see lecture L2) of $\boldsymbol{S}$, that can be reached for a specific value of noise $\boldsymbol{\varepsilon}$:

$$k_r(\boldsymbol{\varepsilon}) \leq \text{cond}(\boldsymbol{S}) = 958 \tag{1.8}$$

3. Example 2: Different inverse problems for steady state 1D heat transfer through a wall

3.1 Case of exact locations

The problem of one-dimension heat transfer through a homogeneous plane wall is considered now. Exact temperature T_e of the $x = e$ rear face is assumed to be known while a sensor located at a depth x_s inside the wall allows the measurement of a temperature y.

Using these two informations and the knowledge of the exact values of the conductivity λ and of the thickness e of the wall, three quantities can be looked for, see figure 2a:

- the temperature T_0, of the other face (x = 0);
- the internal temperature distribution;
- the heat flux density q that flows through the wall.

One temperature is *observed*:

$$T_s = \eta_1 (x_s \,;\, q, T_0, \lambda) \tag{1.9}$$

However, its measurement y by the sensor is supposed to be corrupted by an additive *noise* ε of zero mean and of standard deviation σ:

$$y = T_s + \varepsilon \tag{1.10}$$

The observed temperature T_e can be considered as a particular output of the model η_1 of temperature distribution, at location $x = e$:

$$T_x = \eta_1 (x; q, T_0, \lambda) \equiv T_0 - q\, x / \lambda \tag{1.11}$$

In the parameter estimation terminology:

- T_x is the dependent or output variable,
- x is the explanatory or independent variable,
- q, T_0 and λ are the parameters,
- and function η_1 (**.** ; **...**) is the model structure.

Parameters q, T_0 have a special status: they are also called *input* variables (or *solicitations*), because if they are both equal to zero, the wall temperature field is equal to zero. They correspond respectively to the right hand members of the two boundary conditions of the second and first kinds for the heat equation whose model (1.11) is the solution of what is called a *direct problem*:

$$\frac{\mathrm{d}^2 T}{\mathrm{d}x^2} = 0 \qquad \text{with} \quad -\lambda \left.\frac{\mathrm{d}T}{\mathrm{d}x}\right|_{x=0} = q \qquad \text{and} \quad T\big|_{x=e} = T_e \tag{1.12}$$

We will see later on that this direct problem, whose solution (1.11) is the internal temperature field *in between* the two boundaries ($x = 0$ and $x = e$), is *a well-posed problem*.
The wall conductivity λ is called a *structural* parameter: if its value changes, the material system also changes.

As a consequence of model (1.11), the known value of the rear face temperature verifies:

$$T_e = T_0 - q\, e / \lambda \tag{1.13}$$

Elimination of *q* between the two equations (1.11) and (1.13) yields a second model η_2 for the output of the sensor located in x_s :

$$T_s = \eta_2 (x_s/e, T_0, T_e) = \left(1 - \frac{x_s}{e}\right) T_0 + \frac{x_s}{e} T_e \tag{1.12}$$

Inversion of this second model is straightforward, replacing T_s by its measured value *y*:

$$\hat{T}_0 = \frac{1}{1 - x_s^*}\, y - \frac{x_s^*}{1 - x_s^*}\, T_e \qquad \text{with} \qquad x_s^* = x_s / e \tag{1.13}$$

The hat superscript $\hat{\alpha}$ over a α quantity designates here either an estimator of α, in the statistical sense, that is a random variable whose *realization* is an approximate value of the exact value of α, or its *estimated* (observed) value.

This allows the calculation of the *estimation* error for T_0, $e_{T0} = \hat{T}_0 - T_0$, which is a random variable proportional to ε, of zero mean (symbol E (.) is used here for the mathematical expectancy of a random variable), with its own standard deviation σ_0 :

$$e_{T0} = \varepsilon / (1 - x_s^*) \quad \Rightarrow \quad \mathrm{E}(e_{T0}) = 0 \quad \text{and} \quad \sigma_0 = \sigma / (1 - x_s^*) \tag{1.14}$$

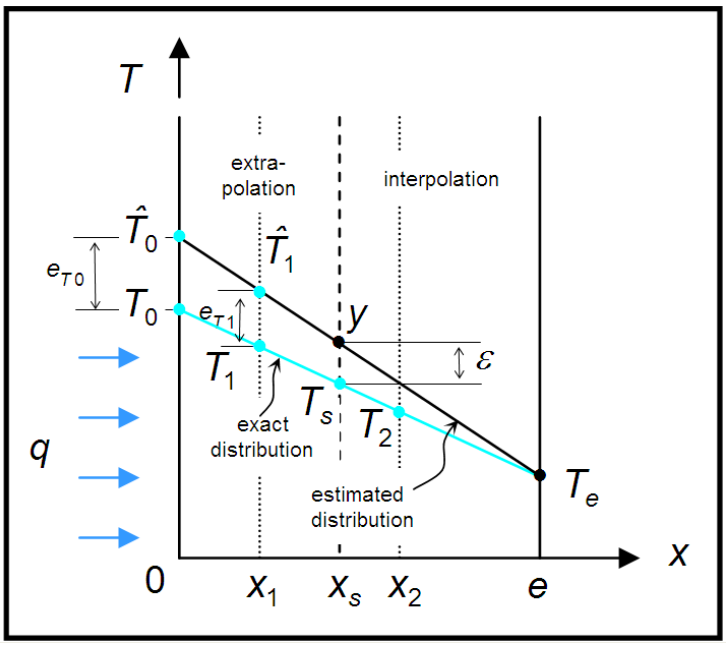


Figure 2a – Estimation of temperature/flux in a wall
Noised temperature measurement
Exact sensor location

A direct consequence of (1.14) is that estimation of T_0 is unbiased, $E(\hat{T}_0) = T_0$, and its standard deviation $\sigma_{T0} = \sigma_0 = \sigma / (1 - x_s^*)$ is an increasing function of the relative depth x_s^* of the sensor inside the wall.

An obvious property of the linear extrapolation related to the straight line model (1.12) can be highlighted:

- error on T_0, measured by its standard deviation σ_0, becomes infinite if the sensor is located at $x = e$ (rear face). It reaches a minimal value for a measurement at the $x = 0$ face;

The estimated temperature distribution that derives from $\hat{T}_0$, also called *recalculated* distribution, is given by $\eta_2(x / e, \hat{T}_0, T_e)$:

$$T_{\text{recalce}}(x) = \eta_2\ (x^*, \hat{T}_0, T_e) = \hat{T}_x = \frac{1 - x^*}{1 - x_s^*}\, y + \frac{x^* - x_s^*}{1 - x_s^*}\, T_e \quad \text{with} \quad x^* = x / e \qquad (1.15)$$

The random error $e_{Tx} = \hat{T}_x - T_x$ for temperature T_x at any depth x, can be assessed by the same type of derivation, as well as its standard deviation σ_{Tx}:

$$e_{Tx} = K\ \varepsilon \qquad \Rightarrow \qquad \sigma_{Tx} = K\ \sigma \qquad \text{with} \quad K = \frac{1 - x^*}{1 - x_s^*} \qquad (1.16)$$

Two regions can be distinguished inside the wall (see figure 2a):

- the external layer, between x_s and e, that is the layer whose points x_2 are located in between boundaries where temperature boundary conditions (1rst kind) are either approximately (y) or exactly (T_e) known: going from y to $\hat{T}_x$ corresponds to a graphical *interpolation* with a reduction of the estimation error with respect to the noise ($K \leq 1$). The inverse temperature T_x estimation problem is *well-posed* in this region.

- layer in between 0 et x_c, with *external* points x_1, where the same operation consists in making an extrapolation. This corresponds therefore to an amplification of the measurement noise ($K \geq 1$) : the inverse problem of estimation of temperature T_x is *ill-posed* in this region.

Remark 5:

This partition of the space domain into two zones, an internal one located between limits where noised boundary conditions are available, and an external one, beyond these limits, leads to ill-posed problems as soon as the temperature field, or its derivative, is looked for in the external zone. This is true not only in this 1D steady state type of diffusion

problem, but also in transient regime, whatever the space dimension (1 to 3D) of the geometrical domain.

An estimation $\hat{q}$ of heat flux q can be given here, as well as an assessment of its error e_q and of its standard deviation σ_q (a statistical quantification of what is called « absolute » error) and of its relative standard deviation σ_q / q (a statistical quantification of what is called « absolute » error):

$$\hat{q} = \lambda \frac{y - T_e}{e - x_s} \Rightarrow e_q = \frac{\lambda}{e - x_s} \varepsilon \Rightarrow \sigma_q = \frac{\lambda}{e - x_s} \sigma \Rightarrow \sigma_q / q = \frac{1}{1 - x_s^*} \frac{1}{SNR} \quad \text{(1.16a, b, c, d)}$$

Let us note that the relative standard deviation of the estimated flux (1.16d) depends on the temperature signal/noise ratio $SNR = (T_0 - T_e)/\sigma$ and on the relative depth x_s^* of the sensor.

We consider a numerical example here. The wall is 0.2 m thick with a thermal conductivity equal to 1 $W.m^{-1}.K^{-1}$, with a 30°C temperature difference between its faces and a 0.3 °C value for the standard deviation of the temperature noise for a measurement in $x_s = 0.18$ m:

$$q = \lambda \frac{T_0 - T_e}{e} = 1 \frac{30}{0.2} = 150 \, \mathrm{W.m^{-2}} \quad \text{and} \quad SNR = (T_0 - T_e)/\sigma = 30 / 0.3 = 100 \quad (1.17)$$

This yields a 10 % error (relative standard deviation) for $\hat{q}$ (see equation 1.16d). A mid-slab measurement ($x_s = 0.1$ m) would have given a 2 % error for this flux: the location of the measurement is therefore a key parameter.

3.2 Case of imprecise sensor locations and errors for parameters "assumed to be known"

Measurement noise is not the only cause of the estimation error: in numerous practical experimental situations, where a sensor has to be embedded in a material, the precise location of its active element (the hot junction of a thermocouple, for example) is not precisely known. So a different type of error has to be taken care of.

Let us assume that, in the above example, the objective is the same (estimation of the front face temperature T_0, of the inner temperature distribution T_x and of the heat flux q), but the sensor which was thought to be positioned at a *nominal* location x_s^{nom} is actually located at depth x_s, with:

$$x_s^{nom} = x_s + \delta \quad (1.18)$$

see figure 2b. So, the noised output *y* of the sensor stems from the error δ in its depth, see figure 2b:

$$y = \eta_2\,(x_s / e, T_0, T_e) + \varepsilon = \eta_2\,(x_s^{nom} / e, T_0, T_e) + \varepsilon' \quad \text{with} \quad \varepsilon' = \delta\,(T_0 - T_e)/e + \varepsilon \tag{1.19}$$

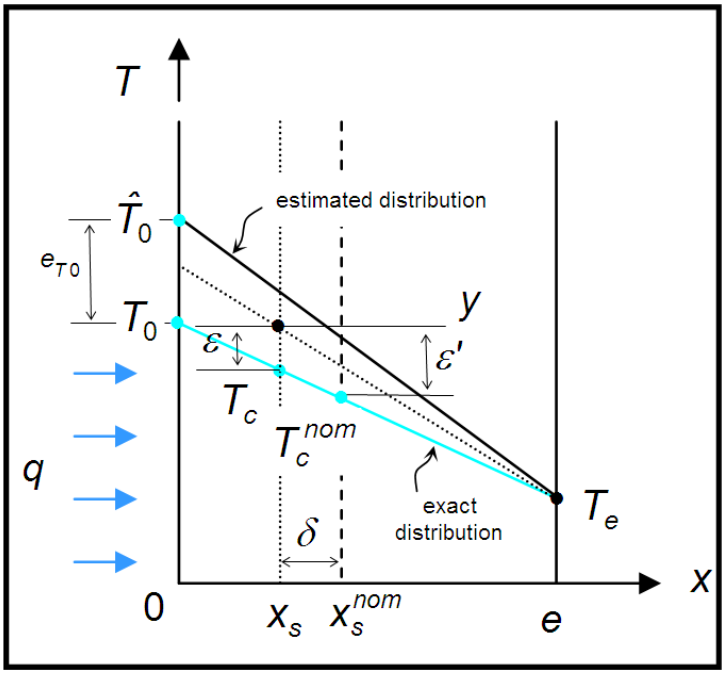


Figure 2b - Estimation of temperature/flux in a wall
Noised temperature measurement
Noised sensor location

If one assumes here that this position error δ is also a random variable, which is independent of temperature noise ε, of zero mean ($E(\delta) = 0$) and of standard deviation σ_{pos}, we find the same type of error as in section 3.1, simply replacing σ by a standard deviation σ':

$$\sigma'^2 = var\,(\varepsilon') = \sigma^2 + \left((T_0 - T_e)/e\right)^2 \sigma_{pos}^2 = \sigma^2\left(1 + SNR^2 / R_{pos}^2\right) \quad \text{with} \quad R_{pos} = e/\sigma_{pos} \tag{1.20}$$

Contribution in σ' of this position error may become important as well as in all the standard deviations of the subsequent estimation errors (σ_{T0}, σ_{Tx} an σ_q) considered in section 3.1, as soon as the signal/position error R_{pos} ratio becomes low with respect to the signal/temperature noise ratio SNR.

Let us go back to the numerical application (1.17), with the additional assumption of a position error of standard deviation 2 mm. These two ratios become:

$$R_{\text{pos}} = e / \sigma_{\text{pos}} = 200/2 = 100 \qquad \text{and} \qquad SNR = (T_c - T_e)/\sigma = 30/0.3 = 100 \tag{1.21}$$

So, in this case, the presence of the position error is equivalent to a 41 % increase of the temperature measurement noise ($\sigma' / \sigma = \sqrt{2}$ here). The consequence would be a 14.1 % error for the estimated flux (for $x_s = 0.18\ \text{m}$).

This problem of *error in the dependent variable* in parameter estimation problems can be solved using *total least squares* [1, 2] or *Bayesian* estimation techniques. The interested reader can also refer to [3, 4, 5].

Let us note that this type of error belongs to a broader class of errors not directly linked to the measurement noise: it concerns the '*parameters supposed to be known*' (but not estimated generally) in a parameter estimation problem.

Such a problem arises if, in the preceding example, thermal conductivity λ is not precisely known. We can assume than a 'nominal' value λ^{nom} is known, but it differs from the exact value λ^{exact} by an error e_λ:

$$\lambda^{nom} = \lambda^{exact} + e_\lambda \tag{1.22}$$

If we refer to the derivations made in section 3.2, this conductivity error will not have any additional effect on the errors on T_0 and T_x. However estimation (1.16) of flux q has to be revisited:

$$\hat{q} = \lambda^{nom}\,\frac{y - T_e}{e - x_s} = \frac{\lambda^{exact} + e_\lambda}{e - x_s}\,(T_s - T_e + \varepsilon) = \frac{\lambda^{exact}\,(T_s - T_e)}{e - x_s}\left(1 + \frac{e_\lambda}{\lambda^{exact}}\right)\left(1 + \frac{\varepsilon}{T_s - T_e}\right) \tag{1.23a}$$

In the case of a small relative error $e_\lambda / \lambda^{exact}$ for the conductivity and for large signal over noise ratio SNR, the preceding equation can be linearized, which yields the relative error e_q / q^{exact} for the estimated flux:

$$q^{exact} + e_q \approx q^{exact}\left(1 + \frac{e_\lambda}{\lambda^{exact}} + \frac{\varepsilon}{T_s - T_e}\right) \quad \Rightarrow \quad \frac{e_q}{q^{exact}} = \frac{e_\lambda}{\lambda^{exact}} + \frac{1}{SNR\ (1 - x_s^*)}\,\frac{\varepsilon}{\sigma} \tag{1.23b}$$

To go further on, it is necessary to assume that λ^{exact} is a random variable of mean equal to λ^{nom} and of standard deviation σ_λ. Taking the variance of equation (1.21b) yields:

$$\frac{\sigma_q}{q^{exact}} \approx \left(\frac{\sigma_\lambda^2}{\left(\lambda^{exact}\right)^2} + \frac{1}{SNR^2\ (1 - x_s^*)^2}\right)^{1/2} \tag{1.23c}$$

If we consider the case given by (1.17) in section 3.1, with $R_{pos} = 0$ (no position error, with $x_s = 0.18$ m), and an error of 10 % for the conductivity, that is e_λ of zero mean around $\lambda^{nom} = 1$ W.m^{-1}.K^{-1}, with a standard deviation $\sigma_\lambda = 0.1$ W.m^{-1}.K^{-1}) the error σ_q / q^{exact} becomes equal to 14.1 % instead of 10 % for an exact conductivity. This error caused by the supposed to be known conductivity can even become dominant error if the sensor is better located ($x_s = 0.10$ m).

The interested reader can refer to lecture L3 in this school to gain a deeper insight onto the effects of the errors on the parameters that cannot be estimated thanks to temperature measurements and that are 'supposed to be known' in thermophysical characterization problems.

4. Example 3: Inverse problem for unsteady state 1D heat transfer through a wall

4.1 Presentation of the direct problem:

We consider a semi-infinite 1D material with constant thermal properties (λ = 43 W.m^{-1}.K^{-1}, a = 1,18.10^{-5} m^2.s^{-1}) submitted to a heat flux depending on time. We can compute the temperature for several depths in the material (z = 0, 1, 1.5, 5, 10mm) by a direct calculation (Finite Element Method, thermal quadrupoles, analytical solution).

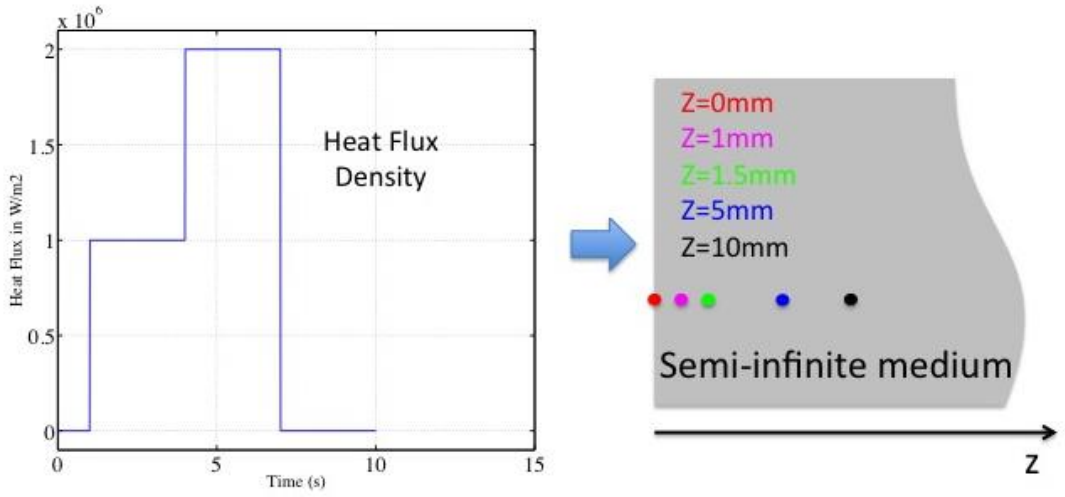


Figure 3a - Heat flux applied to the semi-infinite medium, for several temperature sensor positions

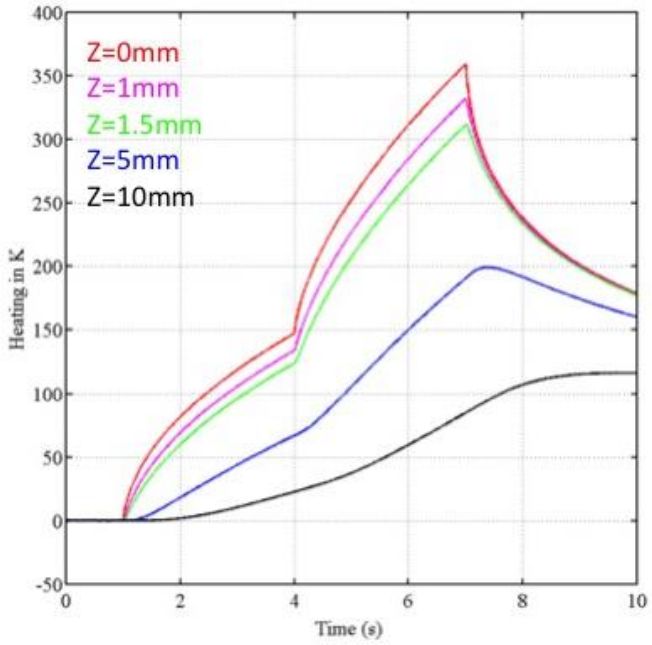


Figure 3b - Corresponding temperature responses for several positions

4.2 Deconvolution procedure, description:

The material is modelled by a linear system subjected to a prescribed heat flux Q (z = 0, t) (noted Q (t) here) having for effect the temperature rise T (z, t). The linear system theory allows to write the temperature T (z, t) as the convolution of Q (t) with the pulse response h (z, t) of the system, (i.e. the material temperature response to a delta function, that is a Dirac distribution, of power density applied to the surface). We assume that the initial temperature distribution in the material (at t = 0) is uniform.

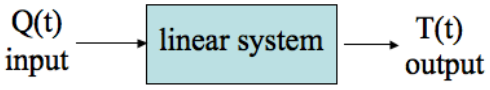


Figure 4 - Linear System

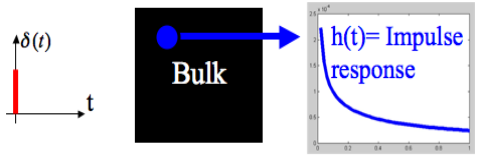


Figure 5 - Impulse response in the bulk

The temperature response T at time t and depth z is:

$$T(z, t) = T(z, t=0) + Q(t)^* h(t) = T(z, t=0) + \int_0^t Q(\tau)\, h(t-\tau)\, d\tau \qquad (1.24)$$

The pulse response h (z, t) of the system is the first time derivative of its step response u(z,t). So, we approximate (1.25) by finite differences which leads to the expression of the

temperature at each time step F in matrix form: where X is a triangular lower square matrix (of order *F*) assembled with the components $\Delta u(z,F) = u(z,F) - u(z,F-1)$ [6]:

$$\begin{bmatrix} \Delta T(z,1) \\ \Delta T(z,2) \\ \vdots \\ \vdots \\ \vdots \\ \Delta T(z,F) \end{bmatrix} = \begin{bmatrix} \Delta u(z,1) & 0 & \cdots & \cdots & \cdots & 0 \\ \Delta u(z,2) & \Delta u(z,1) & \cdots & \cdots & \cdots & 0 \\ \Delta u(z,3) & \Delta u(z,2) & \ddots & \ddots & \ddots & 0 \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & \ddots & \vdots \\ \Delta u(z,F) & \Delta u(z,F-1) & \cdots & \cdots & \ddots & \Delta u(z,1) \end{bmatrix} \cdot \begin{bmatrix} Q(z=0,1) \\ Q(z=0,2) \\ \vdots \\ \vdots \\ \vdots \\ Q(z=0,F) \end{bmatrix} \quad (1.25)$$

$$\Delta \boldsymbol{T} = \boldsymbol{X}.\boldsymbol{Q} \quad (1.26)$$

A noise is added to the numerical signal in order to obtain more realistic data, the new signal can be written (at a given time):

$$\boldsymbol{Y} = \Delta \boldsymbol{T} + \boldsymbol{\varepsilon} \quad (1.27)$$

$\boldsymbol{Y}$ is the new signal. $\Delta \boldsymbol{T}$ is the output of model (1.25), already plotted in figure 3. $\boldsymbol{\varepsilon}$ is a centered zero mean, Gaussian noise with a standard deviation of 0.1 K (it is supposed to be independent). All three preceding quantities are written here in a column-vector form of size (*F* x 1). The deconvolution procedure consists in inverting Eq.(1.26), i.e. expressing surface heat fluxes with measured surface heating:

$$\boldsymbol{Q} = \boldsymbol{X}^{-1}\boldsymbol{Y} \quad (1.28)$$

In the case of the deconvolution of infrared surface temperature (z = 0), the inverse problem is stable and the inversion of matrix $\boldsymbol{X}$ does not cause any problem (see Fig. 6a).

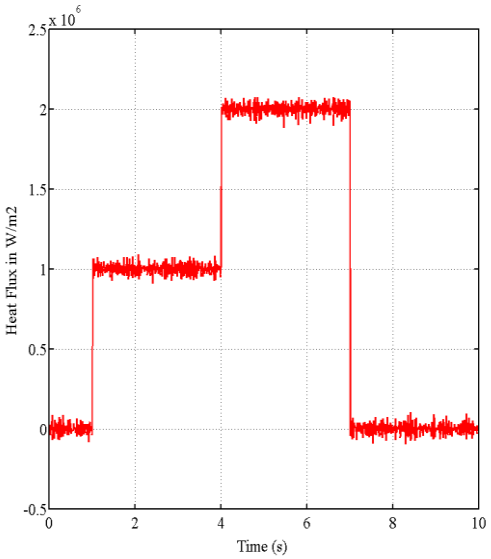


Figure 6a – Estimated heat flux, starting from noisy measurements at z = 0

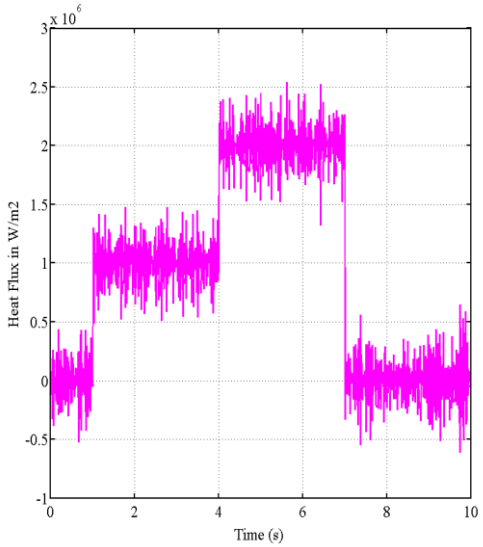


Figure 6b – Estimated heat flux, starting from noisy measurements at z = 1 mm

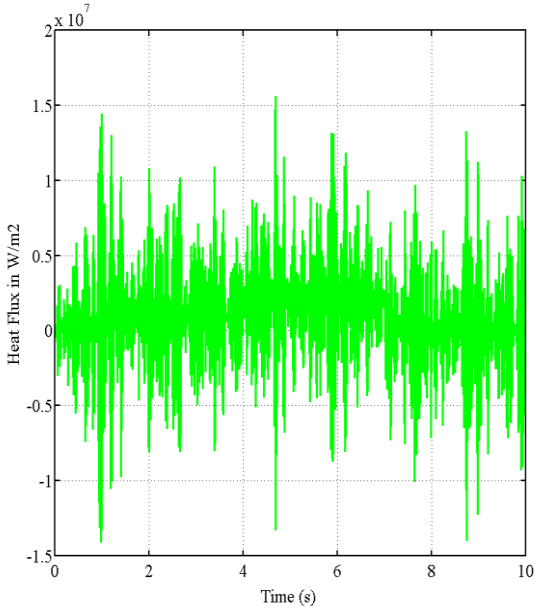


Figure 6c - Estimated heat flux, starting from noisy measurements at z = 1.5 mm

However, in the case of the deconvolution of the temperature measured by a thermocouple, the inverse problem becomes difficult to solve with a good precision because the conditioning of $\boldsymbol{X}$ gets worse. The deeper the thermocouple is located, the more unstable the problem becomes. Clearly, it means that matrix $\boldsymbol{X}$ becomes difficult to invert because of the presence of very small coefficients (in absolute value) in its diagonal: the result does not respect the stability criterion because the noise in $\boldsymbol{Y}$ is amplified. In figure 6b the heat flux estimated with the temperature at z = 1 mm is plotted. The inversion is possible, but the estimated heat flux is very noisy. The heat flux estimated using the temperature at z = 1.5 mm (see Fig. 6c), is too noisy to be exploited: a regularization procedure is needed to find a more stable "quasi solution".

4.3 Regularization procedure

The solution vector $\hat{\boldsymbol{Q}}$, is very sensitive to measurement errors contained in the vector of temperature measurements $\boldsymbol{Y}$. In order to obtain a stable solution, we use a regularization procedure. For example, we can use the Tikhonov regularization operator [7]. The regularized solution becomes:

$$\hat{\boldsymbol{Q}}_{reg} = (\boldsymbol{X}^t\boldsymbol{X} + \gamma \boldsymbol{R}^t\boldsymbol{R})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \tag{1.29}$$

- $\hat{\boldsymbol{Q}}_{reg}$ is the regularized solution (an estimation of $\boldsymbol{Q}$)
- γ is the regularization parameter
- $\boldsymbol{R}$ is the regularization *matrix* depending on the type of information that we want to impose.

In our case, we want a solution with a minimal norm of the solution (0 order) $\left\|\hat{Q}_{reg}\right\|$, so we will take $\boldsymbol{R} = \boldsymbol{I}_d$. An optimal value of the regularization parameter can be found using the "L curve" technique [8]. This type of representation allows to choose the best compromise - which is situated at the bending point of the 'L-curve' - between a stable solution, with a low value of $\left\|\boldsymbol{R}\hat{\boldsymbol{Q}}_{reg}\right\|$ and an accurate solution, with low residuals $\left\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{Q}}_{reg}\right\|$. Another possibility is to use the "discrepancy principle", that is to choose γ such as the root mean square of the residuals gets the same order of magnitude as the measurement noise, that is $\left\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{Q}}_{reg}\right\| \approx \sqrt{m}\ \sigma$, m being the number of measurement times.

Considering the case of the temperature deconvolution at z = 1.5mm (with noise):

- For low values of γ (Fig. 7a.), the solution is unstable with low residuals
- For strong values of γ (Fig. 7b.), the solution is stable but departs from the exact solution.
- For the best compromise of the γ value (Fig. 8) the heat flux is stable and can be used.

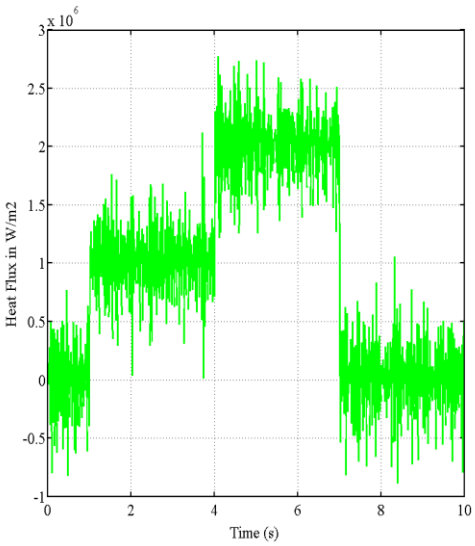


Figure 7a - Heat flux estimation with a low γ

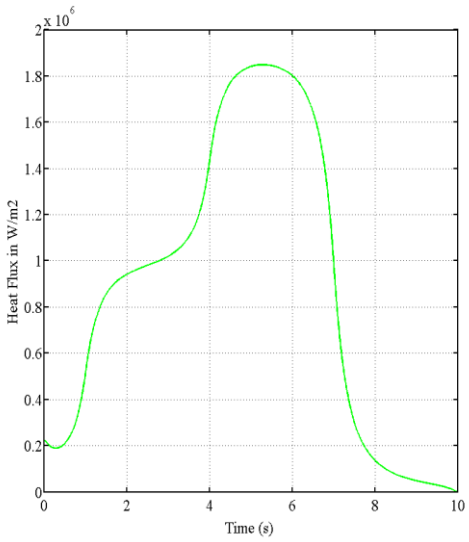


Figure 7b - Heat flux estimation with a large γ.

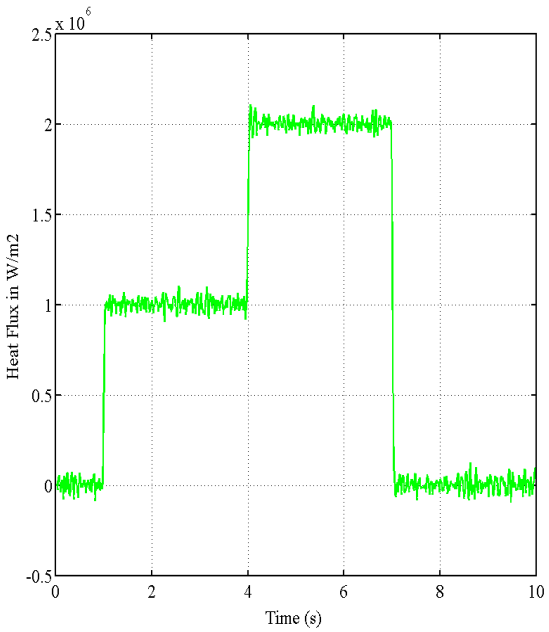


Figure 8 - Heat flux estimation with the best compromise of $\gamma=5.10^{-13}$ $K^2.m^4.W^{-2}$.

One can note that the value γ depends on the level of the noise, the time resolution and the depth of the measurement.

5. Conclusions

The first example presented in this short lecture has been used to precise the notion of an ill-posed problem: under certain circumstances, a small error in the right hand member of a linear system of equations, which can correspond to noised measurements, can yield a very large error in the solution.

Study of the condition number of the corresponding matrix allows to assess the severity of this effect. The reader can refer here to the *Singular Value Decomposition* of this matrix, on which the condition number relies (see further lectures).

In the second example, the inverse 1D steady state input problem has been considered. The very important effect of the location of the temperature sensor on the estimation of the temperature distribution and of the flux through a wall has been highlighted. It has been shown that the temperature noise is not the unique source of error in the estimates.

Errors on the location of the sensor, as well as more generally the effect of the parameters 'supposed to be known', have also to be studied with great care in order to get reliable estimations.

In the third example, the temperature of an "in depth" measurement can be used for a heat flux estimation (an inverse problem of function estimation) depending on time. With a regularization procedure, a quasi solution can be obtained using a regularization parameter depending on the depth of the measurement, the noise, and the time resolution. One can note that the transfer function of the material can be modelled, computed or measured.

References

[1] S. Van Huffel and P. Lemmerling. 2002. *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Dordrecht, The Netherlands: Kluwer, Academic Publishers.

[2] http://en.wikipedia.org/wiki/Total_least_squares

[3] Denis Maillet, Thomas Metzger, Sophie Didierjean, Integrating the error in the independent variable for optimal parameter estimation, Part I : Different estimation strategies on academic cases, e *Inverse Problems in Engineering*, vol. 11, n°3, juin 2003, pp. 175-186.

[4] Thomas Metzger, Sophie Didierjean, Denis Maillet, Integrating the error in the independent variable for optimal parameter estimation, Part II : Implementation to experimental estimation of the thermal dispersion coefficients in porous media with not precisely known thermocouple locations, *Inverse Problems in Engineering*, vol. 11, n°3, juin 2003, pp. 187-200.

[5] Thomas Metzger, Denis Maillet, Multisignal least squares: dispersion, bias, regularization, Chapter 17, *Thermal Measurements and Inverse Techniques*, Editors**:** Helcio R.B. Orlande; Olivier Fudym; Denis Maillet; Renato M. Cotta, Publisher: CRC Press, Taylor & Francis Group, Bocca Raton, USA, 779 pages, May 09, 2011, pp. 599-618.

[6] H.S. Carslaw, J.C. Jaeger, *Conduction of Heat in Solids*, Oxford University Press, 1959.

[7] A.N. Tikhonov, V.Y. Arsenin, *Solutions of Ill-Posed Problems*, V.H. Winston&Sons, Washington, D.C., 1977,

[8] P. Hansen, D. O'Leary, SIAM, The use of the L-curve in the regularization of discrete ill-posed problems, *J. Sci. Comput* 14 (1993) 1487-1503.

Lecture 2: Measurements with contact in heat transfer: principles, implementation and pitfalls

B Garnier[1], F Lanzetta[2]

[1] Laboratoire de Thermique et Energie de Nantes, UMR CNRS 6607, Univ. Nantes, France
E-mail: bertrand.garnier@univ-nantes.fr

[2] FEMTO-ST, Energy Department, Univ. of Franche-Comté, CNRS, Belfort, France
E-mail: francois.lanzetta@univ-fcomte.fr

Abstract. The main objective of this lecture is to make the end users aware of the various physical phenomena and especially of the errors frequently met during temperature and heat flow measurement. The lecture is divided in two main parts dealing with thermal measurement at the macroscale and micro and nanoscales respectively. In Part 1, phenomena that occur in thermometry with contact (thermoelectric effects, thermoresistance) will be presented. For thermometry with contact, the analysis of systematic errors related to the local disturbance of field temperature due to the introduction of sensors will be emphasized. Indeed intrusive effects due to sensors are usually ignored and can be reduced using know-how as will be shown through analytical modeling. Otherwise, the interest in using semi-intrinsic thermocouples will be discussed. The specificities of temperature measurement in fluid flow will be detailed. Finally, heat flow measurement using direct methods (gradient, enthalpic, electric dissipation …) or inverse methods (heat flow sensors with a network of thermocouples) will be reminded.

1. Introduction: General notions about temperature sensors

Mediums are in interaction with the environment, the interaction can be of several types: thermal, electrical, magnetic, liquid or vapor mass transfer, chemical reaction, corrosion … The installation of sensor on or inside the mediums should not modify these interactions. The choice of the sensor is performed so that these interactions do not have an effect on the measurement and on the lifespan of the sensor. For example, a sensor on a surface can modify heat transfer by conduction, convection or radiation. Otherwise, the deposit of a liquid film or a coating modifies emissivity and therefore the radiative heat exchanges. The main consequence is that the temperature provided by the sensor can be very different from the one to measure. One important thing to keep in mind is that temperature measurement is accompanied by parasitic effects which must be well-known.

According to the type of interaction between sensor and medium, one can classify the methods of measurement in three categories:

1. *Methods with direct contact sensor-medium:* in this type of method, the sensor tends to locally equilibrate itself with the medium. If there is perfect adiabaticity of the sensor with the environment, its temperature is equal to that of the medium. However, in thermometric devices, this adiabaticity is usually not perfect.
2. *Methods with contact without physical connection with the environment*: in some cases, the temperature readings are carried out using an optical mean therefore no physical connection exists between the sensor and the environment. In this category, we can find surface temperature

measurement with deposited thermosensitive material such as liquid crystals or photoluminescent salts.

3. *Methods without contact*: in this method, sensors are far from the medium. Despite there is still interactions between them, the sensor is no more in equilibrium with the medium. Such methods are essentially based on radiative heat transfer.

In this lecture, one will discuss temperature measurement with contact. A focus on the main methods (thermoelectric, thermoresistance) will be performed. First of all temperature measurement using thermoelectric effects will be analyzed in various situations (temperature measurement in fluids, in semi-transparent medium, and in opaque medium). The recent progress in thermal measurement at micro and nanoscales using Scanning Thermal Microscopy methods will be presented.

2. Phenomena and sensors for temperature measurement

2.1. Thermoresistances

2.1.1. Metallic probes

They are commonly called Resistance Temperature Detectors (RTD). The thermosensitive parameter in these sensors is the electrical resistance. This one changes according to empirical law such as:

$$R = R_0 [1 + \alpha (T - T_0) + \beta (T - T_0)^2] \quad (2.1)$$

Their respective sensitivities, α, are about 10^{-3} K^{-1} which is rather weak, but their accuracy is rather large and higher than that of the thermocouples (Table 2.1.). In the specified temperature range, their stability is good. The resistor probes have an almost linear answer. A resistance measurement device or a power supply with a low voltage voltmeter has to be used to induce a current of about a few mA through the thermoresistive probes. One has to take care of self-heating or the Joule effect in order to limit temperature bias. For practical applications, the thermoresistive probes are composed of a metallic layer deposited on a flat electrical insulating substrate (epoxy resin, ceramic, mica…) or cylindrical (glass, pyrex….). The size and shape of these thermoresistive probes make them useful for average temperature measurement. In addition, their time constant is much larger than that of thermocouples due to their insulating substrate. Therefore, they will be used preferentially for temperature measurement in stationary mode.

Table 2.1. Characteristics of the main thermoresistive metallic probes

Metal	Sensitivity α (K^{-1})	Temperature range (°C)
Platinum	$4\ 10^{-3}$	-200 à +1000
nickel	$6\ 10^{-3}$	-190 à +350*
Copper	$4\ 10^{-3}$	-190 à +150**

* : 358°C=Curie point for Nickel (magnetic transformation)

** : risk of oxidation for copper

2.1.2. Thermistors

The thermistors which are probes with semiconducting material are much more sensitive than the metallic probes (sensitivity 10 times larger), but they are less stable and their calibration curve is strongly nonlinear:

$$R = R_0 \exp [B (1/T - 1/T_0)] \quad (2.2)$$

The thermistors are presented in several shapes: pearl, disc or rod. The pearls are made of semiconducting material dropped on two connecting wires. Their diameter is about 0.15 to 2.5mm. They can be coated with glass. The flat discs are of more important size (2 to 25 mm in diameter and 0.5 to 12 mm thick). The rods

are metalized at their extremity for the contact with the connecting wires. Their time constant ranges from a few seconds to several tens of seconds and the temperature range for thermistors goes usually from -50°C to 500°C.

2.2. Thermoelectric effects: theory and practice

The thermocouple is the most widely electrical sensor in thermometry and it appears to be the simplest of electrical transducers. Thermocouples are inexpensive, small in size, rugged, and remarkably accurate when used with an understanding of their peculiarities. Accurate temperature measurements are typically important in many scientific fields for the control, the performance and the operation of many engineering processes. A simple thermocouple is a device which converts thermal energy to electric energy. Its operation is based upon the findings of Seebeck [1]. When two different metals *A* and *B* form a closed electric circuit and their junctions are kept at different temperatures T_1 and T_2 (Figure 2.1), a small electric current appears.

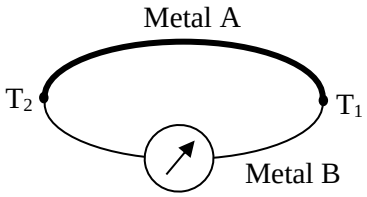


Figure 2.1. Thermocouple circuit.

The electromotive force, emf, produced under these conditions is called the Seebeck emf. The amount of electric energy produced is used to measure temperature. The electromotive force depends on the materials used in the couple and the temperature difference T_1-T_2. The Seebeck effect is actually the combined result of two other phenomena, the Peltier effect [2] and the Thomson effect [3]. Peltier discovered that temperature gradients along conductors in a circuit generate an emf. Thomson observed the existence of an emf due to the contact of two dissimilar metals and related to the junction temperature. The Thomson effect is normally much smaller in magnitude than the Peltier effect and can be minimized and disregarded with proper thermocouple design.

a) Peltier effect

A Peltier electromotive force $V_M - V_N$ is created at the junction of two different materials (wire or film) A and B, at the same temperature *T*, depending on the material and the temperature *T* (Figure 2.2.):

$$V_M - V_N = \Pi_{AB}^T \qquad (2.3)$$

Π_{AB} is the Peltier coefficient at temperature *T*.

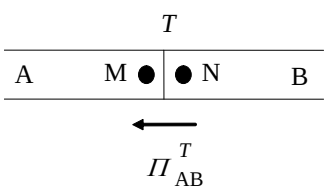


Figure 2.2. Peltier effect without current flow.

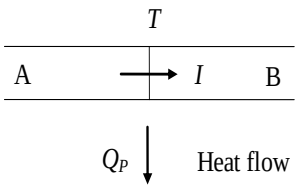


Figure 2.3. Peltier effect with current flow.

When a current *I* flows through a thermocouple junction (Figure 2.3.), heat, Q_P, is either absorbed or dissipated depending on the direction of the current. This effect is independent of Joule heating.

$$dQ_P = (V_M - V_N)Idt = \Pi_{AB}^T Idt \quad (2.4)$$

Q_p is the heat quantity exchanged with the external environment to maintain the junction at the constant temperature T.
The phenomena are reversible, depending on the direction of the current flow and:

$$\Pi_{AB}^T = -\Pi_{BA}^T \quad (2.5)$$

b) Volta's law

In an isothermal circuit composed by different materials, the sum of the Peltier EMFs is null (Figure 2.4.) and:

$$\Pi_{AB} + \Pi_{BC} + \Pi_{CD} + \Pi_{DA} = 0 \quad (2.6)$$

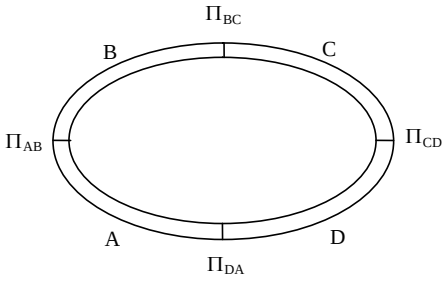


Figure 2.4. Volta's law with four materials.

c) Thomson effect

Thomson EMF's corresponds to the tension $e_A(T_1,T_2)$ between two points M and N of the same conductor, submitted to a temperature gradient, depending only on the nature of the conductor (Figure 2.5.):

$$e_A(T_1,T_2) = \int_{T_1}^{T_2} \tau_A dT \quad (2.7)$$

Where τ_A is the Thomson coefficient of the material A.

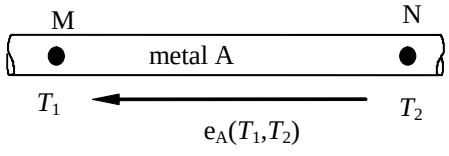


Figure 2.5. Thomson effect without current flow.

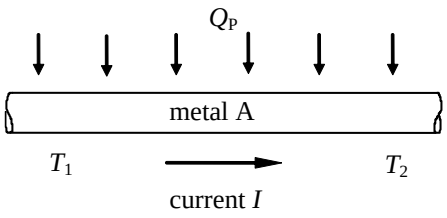


Figure 2.6. Thomson effect with current flow.

When a current I flows through a conductor within a thermal gradient ($T_1 - T_2$), heat Q_T, is either absorbed or dissipated (Figure 2.6.):

$$dQ_T = e_A(T_1,T_2)\, I\, dt = \int_{T_1}^{T_2} \tau_A dT\ I\ dt \quad (2.8)$$

d) Seebeck effect

When a circuit is formed by a junction of two different metals A and B and the junctions are held at two different temperatures, T_1 and T_2, a current I flows in the circuit caused by the difference in temperature between the two junctions (Figure 2.7.).

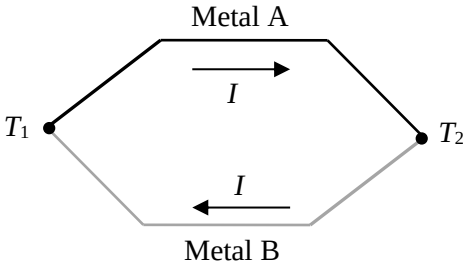


Figure 2.7. Seebeck effect.

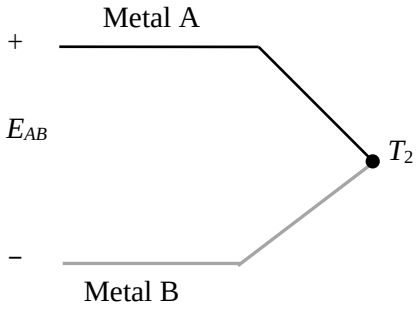


Figure 2.8. Seebeck Voltage.

The sum of the different Peltier and Thomson EMF for the circuit corresponds to the Seebeck EMF:

$$E_{AB}(T_2,T_1) = \Pi_{AB}^{T_1} + \Pi_{BA}^{T_2} + \int_{T_1}^{T_2} \tau_B dT + \int_{T_2}^{T_1} \tau_A dT$$
$$E_{AB}(T_2,T_1) = \Pi_{AB}^{T_1} - \Pi_{AB}^{T_2} + \int_{T_2}^{T_1} (\tau_A - \tau_B) dT \quad (2.9)$$

Then, the Seebeck EMF becomes:

$$E_{AB}(T_1,T_2) = \sigma_{AB}(T_1 - T_2) \quad (2.10)$$

σ_{AB} is the Seebeck coefficient for the A and B metals of the couple ($\mu V.°C^{-1}$ or $\mu V.K^{-1}$). This coefficient corresponds to a constant of proportionnality between the Seebeck voltage and the temperature difference. If the circuit is open at the center of the circuit (Figure 2.8), the net open voltage is a function of the junction temperature and the composition of the two metals.

The thermoelectric power, or sensitivity, of a thermocouple is given by Table 2.2:

$$\sigma_{AB} = \frac{dE_{AB}}{dT} \quad (2.11)$$

Table 2.2. Seebeck coefficients of various thermocouple materials relative to platinum at 0°C [4]

Material	Seebeck coefficient ($\mu V°C^{-1}$)	Material	Seebeck coefficient ($\mu V°C^{-1}$)
Bismuth	-72	Silver	6.5
Constantan	-35	Copper	6.5
Alumel	-17.3	Gold	6.5
Nickel	-15	Tungsten	7.5
Potassium	-9	Cadmium	7.5
Sodium	-2	Iron	18.5
Platinum	**0**	Chromel	21.7
Mercury	0.6	Nichrome	25
Carbon	3	Antimony	47
Aluminium	3.5	Germanium	300
Lead	4	Silicium	440
Tantalum	4.5	Tellurium	500
Rhodium	6	Selenium	900

Thermocouples are made by the association of dissimilar materials producing the biggest possible Seebeck. In industrial processes, the common thermocouples are presented in Table 2.3.

3. Temperature measurement in fluids

3.1. Mathematical modelling

Transient phenomena appear in many industrial processes and many researchers and engineers have been paying attention to the measurement of temperature fluctuations in turbulent reacting flows, compressible flows, boiling, cryogenic apparatus, fire environments, under the condition of simultaneous periodical variations of velocity, flow density, viscosity and thermal conduction in gas [7-14].

Table 2.3. Thermocouple Types [5]

Type	Metal A (+)	Metal B (-)	Temperature range	Seebeck coefficient α (µV/°C) at T°C	Standard error	Minimal error	Comments
B	Platinum-30% Rhodium	Platinum-6% Platinum	0°C to 1820°C	5.96 µV at 600°C	0.5%	0.25%	Idem R type (glass industry)
E	Nickel 10% Chromium	Copper-Nickel alloy (Constantan)	-270°C to 1000°C	58.67 µV at 0°C	1.7% to 0,5%	1% to 0.4%	Interesting sensitivity
J	Iron	Copper-Nickel alloy (Constantan)	-210°C to 1200°C	50.38 µV at 0°C	2.2% to 0.75%	1.1% to 0.4%	For atmosphere reduced (plastic industry)
K	Nickel-Chromium alloy (Chromel)	Nickel-Aluminium alloy (Alumel)	-270°C to 1372°C	39.45 µV at 0°C	2.2% to 0.75%	1.1% to 0.2%	The most widely used because of its wide temperature range, supports an oxidizing atmosphere
N	Nickel-Chromium-Silicium alloy (Nicrosil)	Nickel-Silicium alloy (Nisil)	-270°C to 1300°C	25.93 µV at 0°C	2.2% to 0.75%	1.1% to 0.4%	New combination very stable
R	Platinum-13% Rhodium	Platinum	-50°C to 1768°C	11.36 µV at 600°C	1.5% to 0.25%	0.6% to 0,1%	High temperature applications, resists oxidation
S	Platinum-10% Rhodium	Platinum	-50°C to 1768°C	10.21 µV at 600°C	1.5% to 0.25%	0.6% to 0.1%	Idem R type
T	Copper	Copper-Nickel alloy (Constantan)	-270°C to 400°C	38.75 µV at 0°C	1% to 0.75%	0.5% to 0.4%	Cryogenic applications
W	Tungsten	Tungsten-26% Rhenium	+20°C to +2300°C				Sensitive to oxidizing atmospheres, linear response and good performance in high temperature
W3	Tungsten-3% Rhenium	Tungsten-25% Rhenium	+20°C to +2000°C				Idem W type
W5	Tungsten-5% Rhenium	Tungsten-26% Rhenium	+20°C to +2300°C				Idem W type

There has been considerable progress in recent years in transient thermometry techniques. Some of these techniques are applicable for both solid material characterization while others are suitable only for fluids thermometry. This chapter deals only with temperature thermocouples measurements in fluids (gases and liquids). Many concepts involved in the temperature measurements in fluids are common to both types and they are discussed here. The techniques for temperature measurement in a fluid consist in inserting a thermocouple, allowing it to come to thermal equilibrium, and measuring the generated electrical signal. When a thermocouple is submitted to a rapid temperature change, it will take some time to respond. If the sensor response time is slow in comparison with the rate of change of the measured temperature, then the thermocouple will not be able to faithfully represent the dynamic response of the temperature fluctuations. Then, the problem is to measure the true temperature of the fluid because a thermocouple gives its own

temperature only. The temperature differences between the fluid and the sensor are also influenced by thermal transport processes taking place between the fluid to be measured, the temperature sensor, the environment, and the location of the thermocouple. Consequently, the measured temperature values must be corrected. Whereas in steady conditions only the contributions of the conductive, convective, and radiative heat exchanges with the external medium occur, unsteady behavior introduces another parameter that becomes predominant: the junction thermal lag which is strongly related to its heat capacity and thermal conductivity. The corrections generally decrease with the thermocouple diameters, and both temporal and spatial resolutions are improved. However, while spatial resolution is fairly directly connected with the thermocouple dimensions, the temporal resolution doesn't only depend on the dimensions and the thermocouple's physical characteristics, but also on the rather complex heat balance of the whole thermocouple. To obtain the dynamic characteristics of any temperature probe, we analyze its response to an excitation step from which the corresponding first time constant τ can be defined as :

$$\tau = \frac{\rho\, c\, V}{h\, A} \tag{2.12}$$

τ is the time constant, ρ the density, c the specific heat, V the volume of the thermocouple and A the area of the fluid film surrounding the thermocouple while h is the heat transfer coefficient.

The goal of this work consists in calculating or measuring time constants of thermocouples and comparing their behavior according to different dynamical external heating like convective, radiative and pseudo-conductive excitations.

An accurate calibration method is an essential element of any quantitative thermometry technique and the goal of any measurement is to correctly evaluate the difference between the "true" temperature and the sensor temperature. Figure 2.9. shows the energy balance performed at the butt-welded junction of a thermocouple for a junction element *dx* resulting from the thermal balance between the rate of heat stored by the junction $d\dot{Q}_{th}$ and heat transfer caused by:

- convection in the boundary layer around the thermocouple $d\dot{Q}_{cv}$
- conduction along the wires $d\dot{Q}_{cd}$
- radiation between the wires and the external medium $d\dot{Q}_{rad}$
- contribution of another source of heat power (a laser source in this example) $d\dot{Q}_{ext}$.

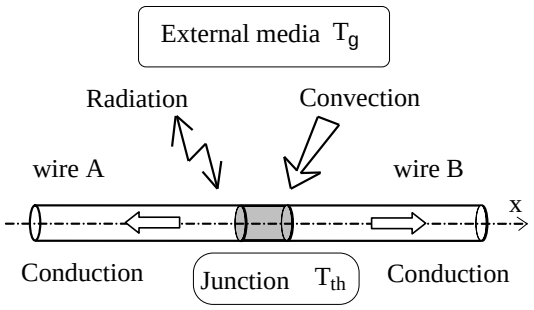


Figure 2.9. Heat balance for the probe

During a transient period, because of its thermal capacity, the thermocouple temperature will lag behind any gas temperature variation. This leads to an error from which a thermocouple time constant can be defined. The general heat balance for a junction of length *dx* is expressed as :

$$d\dot{Q}_{th} = d\dot{Q}_{cv} + d\dot{Q}_{cd} + d\dot{Q}_{rad} + d\dot{Q}_{ext} \tag{2.13}$$

The thermoelectric junction stores the heat by unit time $d\dot{Q}_{th}$:

$$d\dot{Q}_{th} = \rho_{th} c_{th} \frac{\pi d^2}{4} \frac{\partial T_{th}}{\partial t} dx \tag{2.14}$$

where ρ_{th}, c_{th} and T_{th} are the density, the specific heat and the temperature of the junction respectively. The junction is approximated by a cylinder whose diameter equals the wire diameter *d*. This does not exactly fit reality but remains currently used in numerical calculations [15-21]. Moreover, if the wires are uniformly curved, the observation near the junction confirms the previous assumption (Figures 3.20 and 3.21). The Newton's law of cooling is:

$$d\dot{Q}_{cv} = \pi \, dx \, Nu \, \lambda_g \left(T_g - T_{th} \right) \tag{2.15}$$

where λ_g and T_g are the thermal conductivity and the static temperature of the gas. The difficulty is to obtain an accurate relation between the Nusselt number *Nu* and the flow characteristics around the junction assumed as a cylinder [17, 22-25].

Indeed, such a thermocouple is surrounded by both a thermal and aerodynamic gradient which acts as a thermal resistance that is estimated from empiric approaches. A purely convective heat transfer coefficient *h* is generally deduced from correlations about the Nusselt number that is generally expressed as a combination of other dimensionless numbers, such as Eckert, Reynolds, Prandtl or Grashof numbers. However, if many cases have been investigated, the example of thin cylinders cooling process is still an open question. Table 2.4 gives a list of the main Nusselt correlations in this particular case.

Conduction heat transfer $d\dot{Q}_{cd}$ that occurs along the wires to the thermocouple supports has the following general expression:

$$d\dot{Q}_{cd} = \lambda_{th} \frac{\pi d^2}{4} \frac{\partial^2 T_{th}}{\partial x^2} dx \tag{2.16}$$

However, different studies and experiments have shown that conduction dissipation effects along cylindrical wires can be neglected when the aspect ratio between the length and the diameter is large enough [6, 26-32]. Indeed, practical cases of anemometry and thermometry have led to fixing a condition such as:

$$L/d > 100 \tag{2.17}$$

Hence, the temperature gradient can be considered null in the axial direction of the thermocouple wire. The thermocouple is placed in an enclosure at temperature T_w. The enclosure dimensions are assumed to be large with respect to the probe dimensions. Then, the influence of the radiative heat transfer can be expressed in the simplified form:

Table 2.4 Heat transfer laws – These laws describe the heat transfer from a cylinder of infinite length. The film temperature T_{film} is defined as the mean value between the fluid temperature T_f and the thermocouple temperature T_{th} [16-18, 20-25, 29-33]

Author	Temperature for λ, ρ and μ	Correlation	*Reynolds number domain*
Andrews	T_f	$Nu = 0.34 + 0.65\,Re^{0.45}$	$0.015 < Re < 0.20$
Bradley and Mathews	T_f	$Nu = 0.435\,Pr^{0.25} + 0.53\,Pr^{0.33}\,Re^{0.52}$	$0.006 < Re < 0.05$ $0.7 < Pr < 1$
Churchill et Brier	T_f	$Nu = 0.535\,Re^{0.50}\left(T_f / T_{th}\right)^{0.12}$	$300 < Re < 2300$
Collis and Williams	T_{film}	$Nu = \left(0.24 + 0.56\,Re^{0.45}\right)\left(T_{film} / T_{gaz}\right)^{0.17}$	$0.02 < Re < 44$
Collis an Williams	T_{film}	$Nu = \left(0.48\,Re^{0.45}\right)\left(T_{film} / T_{gaz}\right)^{0.17}$	$44 < Re < 140$
Davies and Fisher	T_f	$Nu = \left(2.6/\gamma\pi\right) Re^{0.33}$	$0.01 < Re < 50$
Eckert and Soehngen	/	$Nu = 0.43 + 0.48\,Re^{0.5}$	$1 < Re < 4000$
Glawe and Johnson	T_f	$Nu = 0.428\,Re^{0.50}$	$400 < Re < 3000$
King	T_{film}	$Nu = 0.318 + 0.69\,Re^{0.5}$	$0.55 < Re < 55$
Kramers	T_{film}	$Nu = 0.42\,Pr^{0.2} + 0.57\,Pr^{0.33}\,Re^{0.5}$	$0.01 < Re < 10000$ $0.7 < Pr < 1000$
McAdams	T_{film} and T_f for ρ	$Nu = 0.32 + 0.43\,Re^{0.52}$	$40 < Re < 4000$
Olivari and Carbonaro	T_{film}	$Nu = 0.34 + 0.65\,Re^{0.45}$	$0.015 < Re < 20$ $L / d > 40$
Parnas	T_f	$Nu = 0.823\,Re^{0.5}\left(T_{th} / T_f\right)^{0.085}$	$10 < Re < 60$
Richardson	/	$Nu = 0.3737 + 0.37\,Re^{0.5} + 0.056\,Re^{0.66}$	$1 < Re < 10^5$
Scadron and Warshawski	T_f	$Nu = 0.431\,Re^{0.50}$	$250 < Re < 3000$
Van den Hegge Zijnen	T_{film}	$Nu = 0.38\,Pr^{0.2} + \left(0.56\,Re^{0.5} + 0.01\,Re\right) Pr^{0.33}$	$0.01 < Re < 10^4$

$$d\dot{Q}_{rad} = -\sigma\, \varepsilon(T_{th}) \left(T_{th}^4 - T_w^4\right) dS_{ray} \tag{2.18}$$

σ is the Stefan Boltzmann constant and $\varepsilon(T_{th})$ the emissivity of the wire at the temperature T_{th}. The exchange surface of the radiative heat transfer $dS_{rad} = \pi\, d\, dx$ nearly equals the surface exposed to the convective heat flux. This supposes that the radiative heat transfer between the sensor and the walls is greater than between the gas and the sensor. Here, the assumption is that the gas is transparent, however it is not satisfied in several practical applications like temperature measurements in flames.

In section 3.2.b we will consider a radiative calibration so that the thermocouple junction is submitted to an external heat contribution $d\dot{Q}_{ext}$ produced by a laser beam [27].

$$d\dot{Q}_{ext} = \sqrt{\frac{2}{\pi}} \frac{\left(1-\bar{R}\right)}{a} P_L\, erf\left[\frac{d}{a\sqrt{2}}\right] exp\left[-2\frac{x^2}{a^2}\right] dx \tag{2.19}$$

P_L is the laser beam power, $\bar{R}$ the mean reflection coefficient of the thermocouple junction surface, d the diameter of the junction and a the laser beam radius (this value corresponds to the diameter for which one has *99 %* of the power of the laser beam).

The total heat balance of the thermocouple may be written as follows

$$\begin{aligned} &\rho_{th} c_{th} \frac{\pi d^2}{4} \frac{\partial T_{th}}{\partial t} = Nu\, \lambda_g \pi \left(T_g - T_{th}\right) + \lambda_{th} \frac{\pi\, d^2}{4} \frac{\partial^2 T_{th}}{\partial x^2} \\ &-\sigma\, \varepsilon(T_{th}) \left(T_{th}^4 - T_w^4\right) \pi\, d + \sqrt{\frac{2}{\pi}} \frac{\left(1-\bar{R}\right)}{a} P_L\, erf\left[\frac{d}{a\sqrt{2}}\right] exp\left[-2\frac{x^2}{a^2}\right] \end{aligned} \tag{2.20}$$

The expression of the gas temperature T_g is deduced from equation (2.20):

$$T_g = T_{th} + \tau_{cv} \left[\begin{array}{c} \dfrac{\partial T_{th}}{\partial t} - \dfrac{\lambda_{th}}{\rho_{th} c_{th}} \dfrac{\partial^2 T_{th}}{\partial x^2} + \dfrac{4\, \sigma\, \varepsilon(T_{th})}{\rho_{th} c_{th} d} \left(T_{th}^4 - T_w^4\right) \\ -\dfrac{4}{\rho_{th} c_{th} d^2} \sqrt{\dfrac{2}{\pi}} \dfrac{\left(1-\bar{R}\right)}{a} P_L\, erf\left[\dfrac{d}{a\sqrt{2}}\right] exp\left[-2\dfrac{x^2}{a^2}\right] \end{array} \right] \tag{2.21}$$

Equation 2.21 represents a general expression of the thermocouple dynamic behavior including each of the heat transfer modes. In this expression, the time constant τ_{cv} of the thermocouple junction is defined by:

$$\tau_{cv} = \frac{\rho_{th}\, c_{th}\, d^2}{4\, Nu\, \lambda_g} = \frac{\rho_{th}\, c_{th}\, d}{4\, h} \tag{2.22}$$

If the radiation, the conduction and the external heat supply are neglected, the gas temperature simplifies to:

$$T_g = T_{th} + \tau_{cv} \frac{\partial T_{th}}{\partial t} \tag{2.23}$$

The time-response of a temperature sensor is then characterized by a simple first order equation. This is a common but erroneous way. For a step change in temperature, equation (2.23) reduces to:

$$\frac{T_g - T_{th}}{T_g - T_i} = exp\left[-\frac{t}{\tau_{cv}}\right] \tag{2.24}$$

where T_i is the initial temperature.

Conventionally, the time constant τ_{cv} is defined as the duration required for the sensor to exhibit a *63%*, (= *1-e*$^{-1}$) change from an external temperature step, in the case of a single-order equation. Actually, the fact that different kinds of heat transfers are involved should lead to a global time-constant in which the different phenomena contributions are included [16, 29]. As a consequence, the ability of a thermocouple to follow any modification of its thermal equilibrium is resulting from a multi-ordered time response where the most accessible experimental parameter remains the global time constant. The multi-ordered temperature response of a thermocouple can be represented by the general relation:

$$\frac{T_g - T_{th}}{T_g - T_i} = K_1 \, exp\left[-\frac{t}{\tau_1}\right] - K_2 \, exp\left[-\frac{t}{\tau_2}\right] - \cdots - K_n \, exp\left[-\frac{t}{\tau_n}\right] \qquad (2.25)$$

T_i is the initial temperature, T_g is the fluid temperature. The value of the constants K_1, K_2, … , K_n as well as the time constants τ_1, τ_2, …, τ_n, depend on the heat flow pattern between the thermocouple and the surrounding fluid.

If experiments have shown that most configurations involve nearly first-order behaviors, the measured time-constant does not allow to isolate each of the different contribution modes.

Therefore, the remaining problem of experiments is to relate this global time-constant to the different implied heat transfer modes. Then, our contribution in this section will be to show the influence of the heat transfer condition on the measured time constant value through three different methods of dynamic calibration.

Classical testing of thermocouples often involves plunging them into a water or oil bath and for providing some information only about the response of the thermocouple under those particular conditions. It does not provide information about the sensor response under process operating conditions where the sensor is used. In order to improve thermocouple transient measurements, a better understanding of the dynamic characteristics of the sensor capability is necessary.

3.2. Dynamic calibration

The calibration methods consist of a series of heating and cooling histories performed by submitting the thermocouple to different excitation modes. Then, the resulting exponential rise and decay times of the thermocouple signals allow to estimate the time constant τ. The thermocouple signal is amplified with a low-noise amplifier having a –3 dB bandwidth of 25 kHz (Gain = 1000). The output voltage is finally recorded by a digital oscilloscope.

a) Convective calibration

Figure 2.10. illustrates the convective experimental device. The thermocouple junction is exposed continuously to a constant cold air-stream at constant temperature T_{MIN}. A second hot air flow excites periodically the thermocouple and creates a temperature fluctuation of frequency *f* [33].

The response of a thermocouple submitted to successive steps of heating or cooling is close to a classical exponential first order response from which the time constant can be determined (Figure 2.11.). It can be deduced from the measurement of four temperatures: T_{MAX}, T_{MIN}, $T_{th\,max}$ and $T_{th\,min}$.

For the heating period t_h, we define the temperature differences δ_{1h} and δ_{2h}:

$$\delta_{1h} = T_{MAX} - T_{th\,min} \quad \text{and} \quad \delta_{2h} = T_{MAX} - T_{th\,max} \qquad (2.26), (2.27)$$

For the cooling period t_c, the temperature differences δ_{1c} and δ_{2c} by:

$$\delta_{1c} = T_{th\,max} - T_{MIN} \quad \text{and} \quad \delta_{2c} = T_{th\,min} - T_{MIN} \qquad (2.28), (2.29)$$

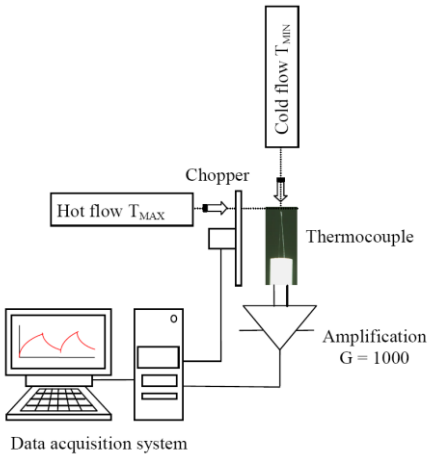


Figure 2.10. Convective characterization setup

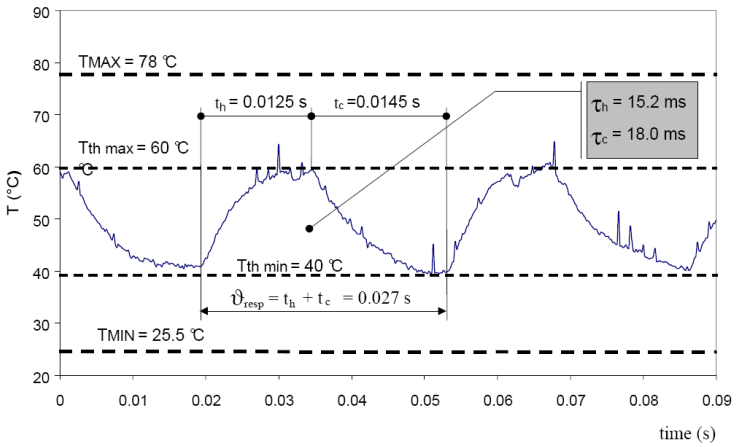


Figure 2.11. Convective characterization results

Then, the two convective time constants are defined while the thermocouple is heating (τ_h) and cooling (τ_c). If we consider a first order response of the sensor we obtain the expressions:

$$\tau_h = \frac{t_h}{ln(\delta_{1h}/\delta_{2h})} \text{ and } \tau_c = \frac{t_c}{ln(\delta_{1c}/\delta_{2c})} \qquad (2.30), (2.31)$$

Then the period of the thermocouple response is:

$$\vartheta_{resp} = t_c + t_h \qquad (2.32)$$

Figure 2.11 presents temperature histories for a 12.7 µm K type thermocouple. The excitation frequency is 37 Hz. The velocities of hot and cold air are both 13 m.s^{-1} at the outlet of the air flow tubes. In any case, the measured time constants are longer during the heating phase than during the cooling one. This phenomenon corresponds to a greater magnitude of the convection coefficient (h). Table 2.5 presents convective time constants for the different thermocouple diameters, resulting from heating periods only and for two air flow velocities (13 m.s^{-1} and 23 m s^{-1}) and for a 5 to 72 Hz explored frequency bandwidth.

Table 2.5 Convective time constant τ_{cv} (ms) and bandwidth Δf (Hz) versus junction diameters. The thermocouple mechanical resistance is not sufficient for the flows with 13 ms^{-1} and 23 ms^{-1} air velocities

Junction diameter		Air velocity : 13 m.s^{-1}		Air velocity : 23 m.s^{-1}	
	d (µm)	τ_{cv} (ms)	Δf (Hz)	τ_{cv} (ms)	Δf (Hz)
S	0.5	–	–	–	–
	1.27	–	–	–	–
	5	2.9	55	2.2	72
K	12.7	15.2	10.5	8.5	18.7
	25	20	8	17	9.4
	250	32	5	25	6.4

One can notice that time constants decrease when increasing the flow velocity because of a larger surface over volume ratio exposed to the flow. Finally, even if the repeatability is good, such a calibration method remains however quite difficult to perform because the fragility of the sensor increases when the wires dimension decreases and the fluid flow increases.

b) Radiative calibration

This calibration method is based on a radiative excitation produced by a continuous argon laser [34, 35]. A set of two spherical lenses allows to locate the beam waist on the junction and an optical chopper generates a periodic modulation of the continuous laser beam. In order to avoid parasitic turbulences around the junction, the sensor is placed in a transparent enclosure (Figure 2.12.).

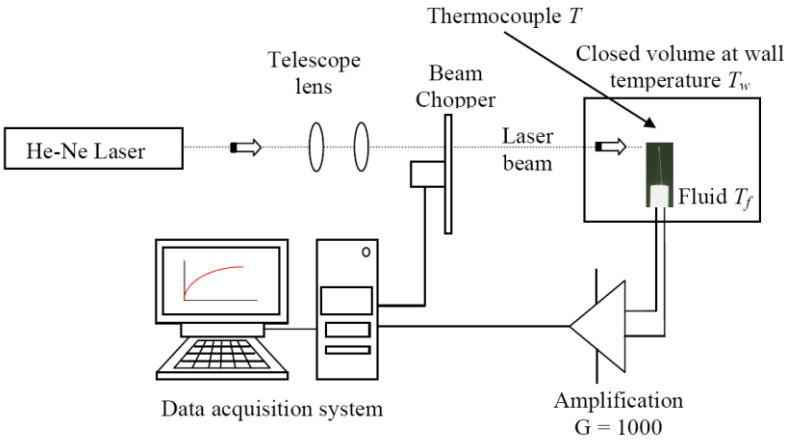


Figure 2.12. Radiative characterization setup

The signal obtained is close to a first order response which gives immediately the sensors dynamic performances. Time constants decreases as diameter and heat transfer (the laser power) increase (Figure 2.13.). This is consistent with the effect of an increasing value of the power density or a decreasing of the beam radius that both acts on the power to heated mass ratio. Table 2.6 presents the radiative time constant for all the thermocouple junction diameters and the explored frequency bandwidth is ranged from 5 to 2274 Hz.

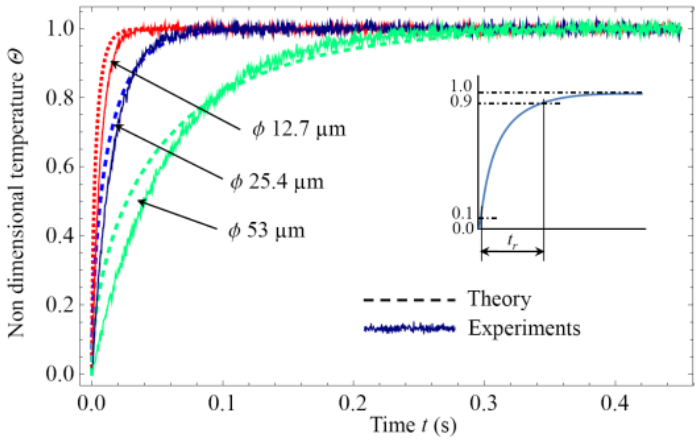


Figure 2.13.

Table 2.6 Radiative time constant τ_{rad} *(ms)* and bandwidth Δf *(Hz)* versus junction diameters

Junction diameter d (μm)		Radiative time constant τ_{rad} (ms)	Bandwidth Δf (Hz)
S	0.5	0.07	2274
	1.27	0.18	884
	5	1.3	123
K	12.7	8.5	19
	25	34	5
	50	64.5	2.5

3.3. Microthermocouple designs

Different methods are used to design a thermocouple probe. It consists of a sensing element assembly, a protecting tube and terminations. Two dissimilar wires are joined at one end to form the measuring junction which can be a bare thermocouple element twisted and welded or butt welded. The protecting tube protects the sensing element assembly from the external atmosphere by a non ceramic insulation, a hard fired ceramic insulator or a sheeted compact ceramic insulator.

The thermocouple probe consists of two wires inserted in a ceramic double bore tube with length and external diameter depending on the experimentation. The wires are cut with a razor blade to produce a flat edge perpendicular to the axis. To realize the junction the thermocouple wires are connected to a bank of condensers (Figure 2.14.).

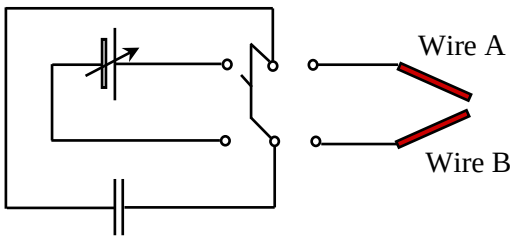
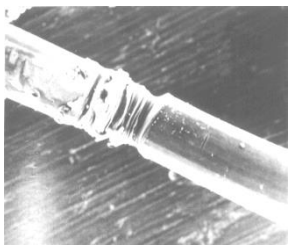


Figure 2.14. Thermocouple welding apparatus



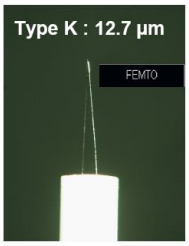


Figure 2.15. Thermocouple junction and probe

The two extremities are approached together in the same time and the beaded junctions are made by a sparking method. The energy release produced by the couple voltage-capacitance is sufficient to weld together the wires. One advantage of this technique is that the resulting junction diameter is not significantly greater than the wires one (Figure 2.15.). Except low mass and specific heat, another consequence is that the cross-sectional area of the wire itself can be used to calculate time constants. A drop of glue can deposited at the tube extremity and pushed down around both wires to minimize the probe fragility.

4. Error introduced by the disturbance of the local temperature using thermocouples

4.1. Introduction

Whatever the selected measurement method, it is accompanied by parasitic effects which must be well-known. The resulting errors can be classified in two categories:

- the ones that are directly related to the thermometric phenomenon, they correspond to the inaccuracy on the measurement of thermometric quantities and to the parasitic effects attached to this phenomenon. It is not here the main topic, but they are not less important. We will quote simply for memory: singularities met in the laws of variation of electrical resistance due to structure modifications (allotropic transformations….), with chemical attacks… and for the thermoelectric circuits, the many parasitic effects such as e.m.f. induced, modifications of the thermoelectric force due to heterogeneities, modifications of structure, junctions nonspecific and not isotherms.

- the others, independently of the selected sensor are related to the fact that the interaction between thermometer, medium and environment causes a local disturbance of the temperature field therefore the local temperature is no more the one that exists before thermometric sensor settling.

In the following, we will present an error analysis and models to describe the local disturbance due to the presence of the sensors. These results come from various works performed at Laboratoire de Thermocinétique, Nantes (Bardon [36], Cassagne [37, 38])

4.2 Error analysis and model

4.2.1 Surface temperature measurement

The surface heat exchanges are modified by the presence of the sensor which does not have the same thermophysical and radiative properties and the same convective heat transfer as the medium to which it is applied. Therefore, a parasitic heat flow is transferred from the medium towards the sensor then from the sensor towards the environment as illustrated in figure 2.18. for surface temperature measurement.

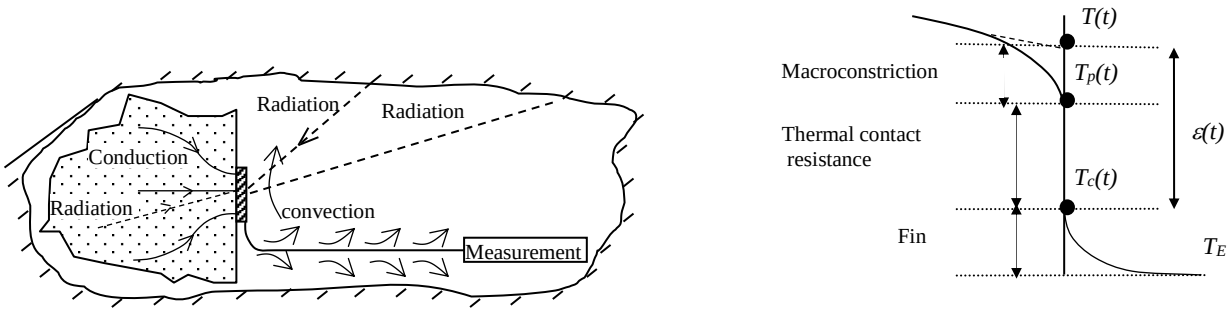


Figure 2.18. Surface temperature measurement

A heat generation or absorption closed to the sensor or to its connection can also occur. All these transfers induce, at the measurement location, a local temperature disturbance which can be either positive or negative according to the heat direction (going in or out). The temperature is no more T but T_p. Moreover, the sensor temperature is not usually equal to T_p because the imperfect contact conditions between sensor and medium involves a temperature discrepancy T_p-T_c which increases as the thermal contact resistance or the heat flux increases.

For an opaque medium, the following three effects are combined:

1. the effect of convergence of heat flux lines towards the sensor (macroconstriction effect),
2. the effect of thermal contact resistance which involves a temperature jump at the sensor/medium interface, and
3. the fin effect which corresponds to the heat transfer towards the outside (over the sensor and along its connection wires).

The measurement error is then:

$$\varepsilon(t) = T(t) - T_C(t) \qquad (2.33)$$

4.2.2. Temperature measurement within a volume

For temperature measurement within a volume, the analysis is similar to the previous one. The error independently of the chosen sensor depends on the fact that the sensor temperature almost never coincides with that of the small element which it replaces. The thermophysical characteristics of the sensors (λ, ρ, c) and its radiative properties are different from those of the medium.

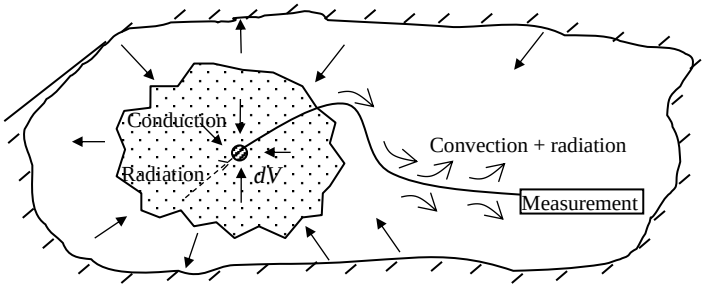


Figure 2.19. Temperature measurement within a volume

Heat transfer within the medium is modified by the presence of the sensor and similarly to surface temperature measurement, a local disturbance of the temperature field appears due to the heat transfer from the medium to the outside through the sensor. One still finds the three effects of: 1) convergence of the heat flux lines towards the sensor. 2) the thermal contact resistance effect 3) the fin effect. In addition the error is still : $\varepsilon(t) = T(t) - T_C(t)$

4.2.3. Error model

The study of the error related to the disturbance of the local temperature requires the solution of a muldidimensional heat transfer problem with various possible configurations and boundary conditions. In this section, one will use relatively simple but very typical models that will clearly show the respective role of conduction within the medium, of non perfect contact between sensor and medium and finally the heat exchanges towards the environment. Most of the conclusions could be extended to numerous others configurations.

We will suppose that the heat exchanges of the medium or of the thermometric connection with the environment can be represented by the heat transfer coefficient, h, and the outside equivalent temperature, T_E. It is known, for example, that for a surface that absorbs a heat flow F (radiation coming from a high-temperature heat source) which exchanges by convection with a fluid at T_f temperature and by radiation with walls at temperature T_0, one has:

$$h = h_c + h_r \qquad (2.34)$$

$$T_E = \frac{h_c T_f + h_r T_0 F}{h} \qquad (2.35)$$

where h_c is the convection heat transfer coefficient, $h_r = 4A\sigma T_m^3$ the radiation coefficient (A is a coefficient which depends on the emissivity and of the relative location of surfaces between which the radiative heat exchange occurs, T_m is an intermediate temperature between T_0 and that of the surface).

4.2.3.1. Steady state surface temperature measurement of an opaque medium

One will investigate surface temperature measurement on an opaque medium of thermal conductivity λ with a simplified sensor having the shape of a rod perpendicular to the surface (figure 2.20.). Far from the sensor, the medium is at the constant temperature T. The surface of the medium is assumed adiabatic except at the contact area S with the sensor.

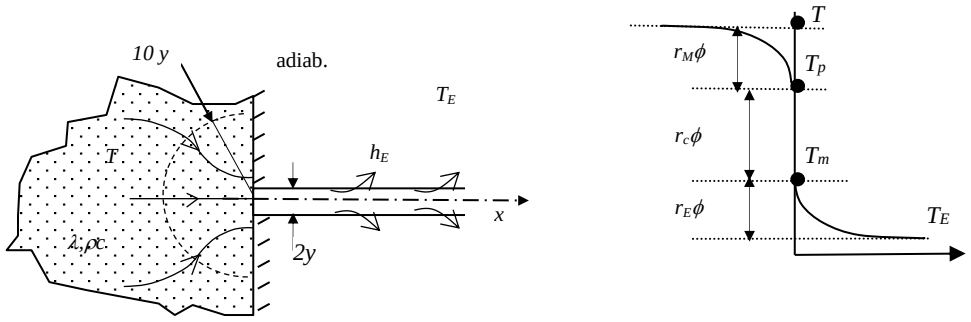


Figure 2.20. Steady state configuration

The three following effects occur due to heat leakage through the sensor towards the outside :

a) The convergence effect : it results from the relation between true temperature and disturbed temperature:

$T - Tp = r_M \phi$ (2.36)

where r_M is a macroconstriction resistance and ϕ the parasitic heat flux . With 3D heat transfer calculation, one can show that $r_M \cong \frac{0{,}4789}{\lambda\sqrt{s}}$ and for a circular surface of radius ρ : $r_M = \frac{1}{4\rho_o\lambda}$.

It is also shown that 96% of the $T - T_p$ temperature drop is within an hemisphere of center 0 and radius 10ρ or $5{,}7\sqrt{s}$.

b) *The contact resistance effect* : responsible for the $T_p - T_c$ temperature drop, it is expressed by :

$T_p - T_c = r_c \phi$ (2.37)

where r_c represents the thermal contact resistance for the surface S (if R_c is the resistance per unit of surface: $r_c = R_c/S$). This effect is related to the imperfection of the contact which results from the irregularities of surfaces. The contact between two solid media is carried out only in some areas (~ 1% of the apparent surface) between which remains an interstitial medium.

c) *The fin effect:* It is responsible for the heat transfer between the connection of the sensor and the environment. Whatever the assumed shape of the connection (rods with uniform or variable section) the heat flux ϕ transfered from the face at $x = 0$ to the environment is linked to the temperature difference (between T_c at x = 0 and the equivalent outside temperature T_E)defined by:

$T_c - T_E = R_E \phi$ (2.38)

where T_c is the temperature at $x = 0$, T_E the equivalent outside temperature and R_E the total thermal resistance between the face x = 0 and the environment. It depends in particular on the geometry, the heat transfer coefficient and the thermal conductivity λ_E of this external connection :

$R_E = 1/(\pi y_E \sqrt{2h_E \lambda_E y_E})$ for a thermocouple assumed as a rod of radius y_E . From relations (2.36, 2.37 and 2.38), one can deduce the heat flux: $\phi = \dfrac{T - T_E}{r_M + r_c + r_E}$ and the measurement error :

$$\delta T = K (T - T_E) \qquad (2.39)$$

with

$$K = \frac{1}{1 + \dfrac{r_E}{r_c + r_M}} \qquad (2.40)$$

The error is thus proportional to the measured and equivalent outside temperatures difference (T-T_E) , the "error coefficient " K is all the more small as the sum of resistances of macro-constriction r_M and contact r_c will be small compared to the external resistance r_E . Therefore, it results that:

- For measurements on a high thermal conductivity medium (metal) , $r_M << r_c$, the thermal contact conditions determines the errors
- For measurements on a low dielectric material, $r_M >> r_c$, the effect of macroconstriction determines the error.
- The roles of r_E and T_E are finally very important. One needs the largest possible r_E and T_E nearest to T (probe with heat flux compensation). It is worth to focus one's attention to the heat flux ϕ_E generated on the surface of the connection wire. If T_E can becomes much higher than T, the error is changed by sign and is of great amplitude: it is necessary to avoid the external radiation of source on the connection. These conclusions, found for the temperature measurement on an opaque medium and for a simplified configuration of a sensor having the shape of a rod perpendicular to surface, remain valid with slight differences for real configurations.

4.2.3.2 Transient surface temperature measurement of an opaque medium

For a fast sudden contact between an opaque medium and a sensor assumed as a rod and perpendicular to its surface, the error becomes function of time: $\varepsilon(t) = K(t)\,[T(t) - T_E]$. It remains proportional to the temperature difference: $T(t) - T_E$

The coefficient $K(t)$ is maximum for $t \to 0$ and decreases for higher t values. For $t \to \infty$, one have: $K(T) \to K(\infty)$ which is obtained for a steady state. The contact conditions between sensor and medium are of great importance :

- If $r_c \neq 0$, $K(0)=1$, the error is about 100% at $t=0$ and decreases all the more the contact between sensor and medium is good.
- If $r_c = 0$ (perfect contact), the initial error is smaller:

  $K(0) = \dfrac{b}{b + b_E} < 1$ where $b = \sqrt{\lambda \rho c}$ and $b_E = \sqrt{\lambda_E \rho_E c_E}$ are the medium and connection effusivities

One can characterize the thermal inertia by the time response x %, such as (figure 2.21.): $\dfrac{K(t_x) - K(\infty)}{K(0) - K(\infty)} = x$

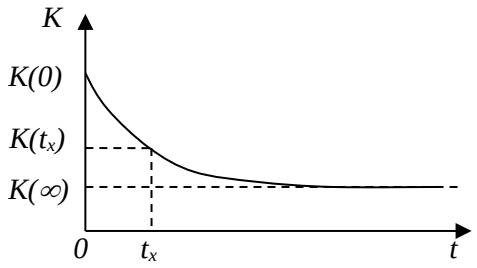


Figure 2.21.

For the same sensor t_x depends strongly on the characteristics of the medium and of the connection medium/sensor/environment. For a conducting medium, t_x depends strongly on r_c which appears as the main factor that determines the sensor inertia. t_x decreases when r_c decreases. It is the same thing if the diameter of the connections is reduced.

For fast transient evolutions, it is worth to weld wires on the surface, so that $r_c \rightarrow 0$, and to use wires as thin as possible. In this case $(r_c \sim 0)$, the thermal inertia t_x is primarily determined by the establishment time t^* of the macro-constriction phenomena within the medium. In practice, this phenomenon remains extremely localized within the immediate vicinity of the sensor (hemisphere of radius $10y$), one can deduce an order of magnitude for t^* by considering the characteristic time $t^* \approx 100y^2/a$ associated to this hemisphere. %. One can consider that, at this time t^*, constriction is established at 97%. One can thus consider that $t_x \approx t^* \approx 100y^2/a$. For temperature with insulating mediums, r_c does not have any effect but t_x is much higher. For a transient evolution with a characteristic time t_c it is worth to choose a sensor for which $t_x << t_c$. In this case, as soon as $t > t_x$, the error will reach, at every moment, its minimal asymptotic value and the steady error model (K_∞) could be applied.

4.2.3.3. Temperature measurement within a volume

In this case, the connection wires usually do not follow an isothermal path on a sufficient length, therefore heat leakage through the sensors occurs. Measurements within a volume are in general much easier than on a surface and errors are usually smaller. However their analysis is more difficult to carry out especially because of the interaction between the connection wire and the medium. In addition, a cavity has to be realized for sensor introduction.

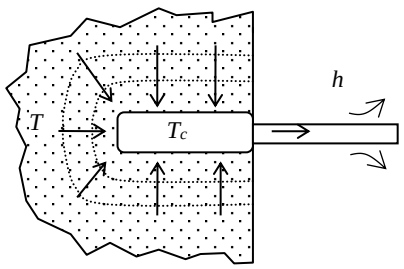


Figure 2.22. Temperature measurement with a cylindrical sensor inside the medium

Therefore, the cavity and sensors don't match exactly, so there exists, between them, some residual space filled with air, grease, glue… which introduces a thermal resistance between sensor and medium. The measurement errors introduced by these phenomena are qualitatively rather similar to those described for surface temperature measurements. Lastly, for long enough isothermal path, heat transfer between sensor and environment is negligible, the differences between thermophysical characteristics (conductivity, heat capacity) of the medium, of the probe or the wire of connection or residual space,

introduce a localized disturbance of the thermal field, and a measurement error remains, but this one is much smaller. An example is provided in figure 2.22.

With this configuration the previous error model (2.39 and 2.40) is still valid, the value of r_c and r_M being different. If we consider that the sensitive element of the sensor with a length L, and a radius y, which recovers its surface S, is isothermal and that it temperature is T_c (figure 2.22.). The contact between the probe and the medium is supposed to be imperfect, therefore for the whole surface of the sensor, the thermal contact resistance r_c is : $r_c = R_c/S$ with $S=2\pi yL$. If $T_E \neq T$, a heat flux occurs between the medium and the environment. The temperature field is modified. In this case the thermal constriction resistance is expressed by: $r_M = \frac{1}{2\pi\lambda\ell} Log \frac{2\ell}{y}$ if $\ell >> y$

4.3. Practical consequence and examples, semi intrinsic thermocouples

4.3.1. Practical consequences

The steady state error model for the simple configuration allows some important features, most of them being valid for other configurations:

1) first of all even for perfect contact $r_c = 0$ there is an error which depends on the ratio r_M / r_E.
2) if the medium is a *high thermal conductivity* material, the macro-constriction r_M will be usually small relatively to r_c and the error will be especially determined by r_c. Thus, one must take care that r_c is small and remains stable. The contact pressure will have to be high and constant, surface will have to be plane without waviness, the interstitial medium with the highest possible thermal conductivity (welding, grease…). In addition, one should avoid oxide films as well as mechanical shocks and vibrations which can modify considerably r_c and consequently the measurement error.
3) For measurements on *an insulator*, r_M is large, usually much higher than r_c. Thus, the macro-convergence effect is the main factor in the measurement error and one can reduce it by increasing the radius of the sensitive element without increasing the section of the connections (figure 2.23.). A contact disc of high thermal conductivity material will be used.

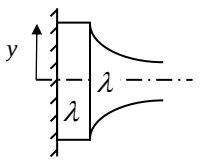


Figure 2.23.

4) Whatever the type of measurement, the fin resistance r_E should be as high as possible. The transversal area, the conductivity, the heat transfer coefficient have to be chosen the smallest possible. One also should have low emissivity surface, connection protected from high temperature fluids movements or radiation, T_E being modified in those situations. One should note that having an insulating layer on the metallic wire of the thermocouple can increase the side heat transfer and therefore the measurement error.
5) Finally, the error is all the more small as T_E is close to the temperature to measure T. It changes with T_E. At the price of a technological complication, one can add an external heat source on the connection so that its temperature T_E is controlled in order to stay a close as possible as T. In this case, one reduces considerably the heat transfer and consequently the error of measurement. This principle is well known as "compensated heat flux sensors". However for correct measurement, the thermal resistance r_E should stay high in order to prevent the compensation heating from disturbing the temperature field in the medium.

4.3.2. Application -for steady state temperature measurement for a thermocouple with and without a contact disc

The two thermocouple wires are considered as a unique rod with a radius y_B = 0.5 mm, an infinite length, an average thermal conductivity λ_B = 15 W.m^{-1}.K^{-1} and a heat transfer coefficient $h_B = 5\ W.m^{-2}K^{-1}$.

The fin thermal resistance is: $r_B = \frac{1}{\pi y_B \sqrt{2h_B y_B \lambda_B}}$ (rod approximation)

Thus, the connection resistance is:

- $r_E = r_B$ without contact disc,
- $r_E \approx r_B + \frac{1}{4y_B\lambda_D}$ with contact disc

($\frac{1}{4y_B\lambda_D}$ is the resistance due to heat flux convergence from y to y_B inside the sensor).

Table 2.9. provides the values of r_M, r_c, r_E and K and for various λ_D with and without disc ($y=y_B= 10$ *mm*, $\lambda_D = \lambda_B$) and for different values of R_c per unit of area:

Table 2.9. Effect of medium thermal conductivity and of the disc on r_M, r_c, r_E and K

	Low thermal conductivity $\lambda=10^{-1}\ W.m^{-1}.K^{-1}$		High thermal conductivity $\lambda=100\ W.m^{-1}.K^{-1}$	
	without disc	with disc	without disc	with disc
$r_M\ (K.W^{-1})$	5000	250	5	0.25
$R_C\ (K.W^{-1}m^2)$	10^{-3}	10^{-3}	10^{-4}	10^{-4}
$r_C(K.W^{-1})$	1270	3,18	125	0,31
$r_E\ (K.W^{-1})$	1700	1733	1700	1733
K	0.786	0.127	0.072	0.0003

5. Temperature measurement with semi intrinsic thermocouple

In this device, one uses the medium M itself (presumately electrically conducting) as one item of the thermocouple (figure 2.24.). Compared to a traditional sensor, this device has several advantages:

- it has only one connection wire instead of two, thus heat leakage is reduced and the thermal resistance r_E is twice larger.
- the measured temperature T_μ is intermediate between T_p and T_c (figure 2.24.)

Figure 2.24. Semi intrinsic thermocouple

For times $t > t_c$ (time constant), T_μ is such that: $\frac{T_p - T_\mu}{T_\mu - T_C} = \frac{\lambda_A}{\lambda_M}$ (2.41)

The error $\varepsilon_i = T - T_\mu$ is thus lower and the contact resistance effect is partly cancelled. For steady state, the error is such that:

$$\varepsilon_i = K_i \left(T - T_E \right) \quad (2.42)$$

With $K_i = \frac{r_M + r_c \frac{\lambda_A}{\lambda_A + \lambda_M}}{r_M + r_c + r_E}$ (2.43)

This error is considerably lower than with a traditional thermocouple (2 to 5 times) and this as much more as the wire thermal conductivity λ_A is small compared to λ_M. In transient mode, error and thermal inertia are greatly reduced (Bardon [36], Cassagne [37]). However, the calibration of the semi intrinsic thermocouple is almost always required. It is usually performed by comparison with a traditional thermocouple.

6. Heat flux measurement: direct and in direct methods

6.1. Direct measurement

6.1.1. Heat flux sensor with gradient (Ravaltera [39])

The principle of this heat flux measurement consists in directly applying the Fourier's law by measuring a temperature difference within the wall itself (intrinsic method) or by covering it with an additional wall (heat flux sensor-HFS-). The surface characteristics of this HFS should be close to those of the wall. The wall of the HFS can be homogeneous (the temperature difference is measured between its two main faces -normal gradient heat flux sensor - figure 2.25.-) or it can be heterogeneous creating heterogeneous temperature that is measured (tangential gradient heat flux sensor – figure 2.26.-).

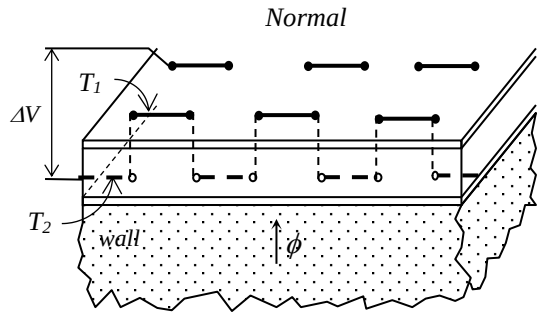


Figure 2.25.

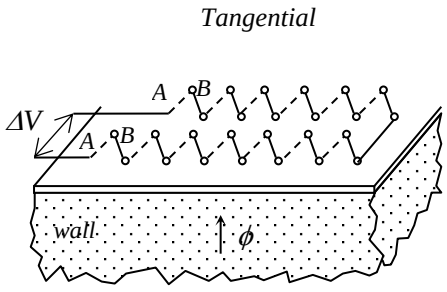


Figure 2.26.

The installation of such HFS on a wall, more or less disturbs the heat flux which crosses it. All must be done so that internal and contact thermal resistances are minimal. In these devices, the measurement of the temperature difference is performed using several thin film thermocouples or thermoresistances. These HFS can work whatever the heat flux direction in steady state or for slowly variable temperature.

Table 2-10 shows a list of commercially available normal or tangential heat flux sensors

Table 2-10 : Main characteristics of available commercial normal and tangential heat flux sensors

Heat flux sensor	**N /T**	**Dim [mm]**	**T_{max} [°C]**	**Sensitivity [μV/(W.m^{-2})]**	**Thick [mm]**	**R_{th} x 10^3 [m^{2}K/W]**	**Response time [s]**	**Accuracy Heat flux [%]**
Hioki	N	10 x10	150	13	0.28	1.4		
RdF/Omega	N	11.93 x25.4	150	1.0	0.125	0.88	0.09	~8-10
RdF/omega	N	11.93 x25.4	150	3.48	0.33	2.1	0.4	~8-10
Wuntronic FM120	N	7.4 x10.7	150	2.64	1.5	4.75	3	5 (Fab)
Flux Teq	N	25.4 x25.4	120	0.8	0.38	0.65	0.6	
GreenTEG gskin	N	10 x10	150	50	0.5	0.35	0.2/0.7	3
Captec	T	10 x10	120	3 à 5		1	0.3	~8-12

N: normal; T: tangential

6.1.2. Inertia heat flux sensor and heat flux sensor with electric dissipation (zero method)

Inertia heat flux sensors works only for variable temperature and if the heat flux is received by the wall. The HFS replaces a piece of the wall and is isolated from this one. Its surface characteristics are identical to those of the wall. The temperature increase of the HFS is proportional to the absorb heat flux and inversely proportional to its capacity (figure 2.27.). The choice of this one is very important because it determines the measurement sensitivity.

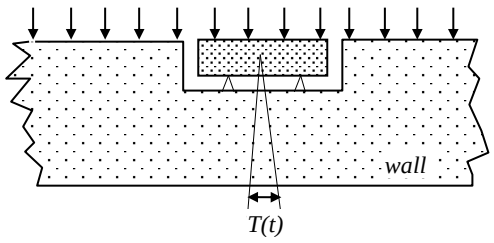


Figure 2.27.

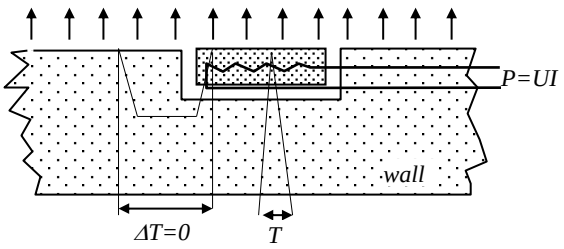


Figure 2.28.

The principle of this HFS with electric dissipation consists in substituting a piece of the wall at its surface with a small heating part insulated towards the wall (fig. 2.28.). The electric heating output is adjusted so that the surface temperature of the wall and of the heating part are equal (*ΔT=0*). Thus, the dissipated electric flux is equal to the heat flux which leaves the wall in its immediate vicinity. This HFS works only for heat flux leaving the wall and for steady state or slowly variable temperature.

6.1.3. Enthalpic heat flux sensor

They are used to measure the heat flux coming from the outside. The HFS replaces an element of surface of the wall and is insulated from this one (figure 2.29.). An initially temperature controlled fluid circulation is heated by the heat flux which induces an enthalpic flow rate. For a correct measurement, the fluid temperature must be adjusted so that wall and HFS temperatures are almost equal. This condition is not always realized and can be an important source of error. The choice of the heat-storage capacity of the fluid also is important..

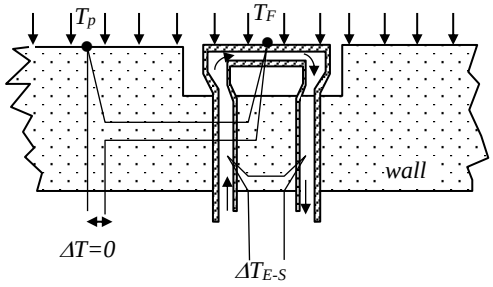


Figure 2.29.

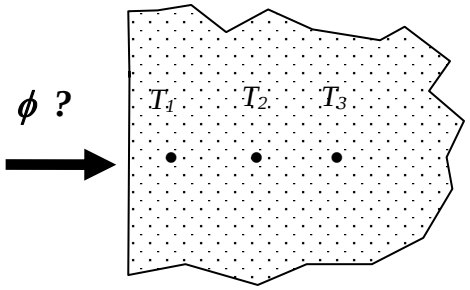


Figure 2.30.

6.1.4. Indirect measurement

One can obtained the surface characteristics (temperature *T*, heat flux φ) from measurements realized within the medium and using inverse methods (figure 2.30.). This procedure involves solution of ill-posed problems. Indeed, one cannot insure a solution, its uniqueness or stability. To solve such difficulties, the technique consists in replacing the ill posed problem by a well posed approximate problem. The solution is found by minimizing a norm of least square type. A heat transfer model (analytical or numerical) is required to solve the direct problem at each optimization step. These methods require significant developments (Beck [40], Alifanov [41], Ozisik [42], Jarny [43]). They will not be presented here. We will just underline that the solution of the inverse problem allows to compute the temperature residuals between final and measured temperatures. These residuals are of great importance because they allow to check the validity of the chosen heat transfer model. If no signature is observed (the residuals are purely random) the model is correct, otherwise the model should be improved.

With regard to the theoretical aspects of the instrumentation, Bourouga [44] has proposed criteria for correct locations of thermocouples to obtain unbiased results and also to optimize the experiment for wall heat flux or temperature estimation.

7. Conclusion

Accurate temperature measurement is not an easy task. Errors depend on thermosensitive phenomena and also according to the sensors which can create local temperature disturbance and therefore bias. Very often, this latter error is ignored. In this lecture dedicated to contact temperature measurement, one have tried to provide to the readers the know-how in various situations (temperature measurement in fluids or opaque medium) in order to perform the best temperature measurements as possible.

8. References

[1] Seebeck, T.J. 1823. Magnetische polarisation der metalle und erze durch temperatur-differenz, *Abh. K. Akad. Wiss.* Berlin, 265

[2] Peltier, J. C. A. 1834. *Annales de Chimie et de Physique* 56: 371-470.

[3] Thomson, W. 1848. On an Absolute Thermometric Scale founded on Carnot's Theory of the Motive Power of Heat, and calculated from Regnault's Observations, *Philosophical Magazine* 33:313-317.

[4] Rathakrishnan, E. 2007. *Instrumentation, Measurements and Experiments in Fluids*, CRC Press.

[5] Devin E. 1997. Couples thermoélectriques, données numériques d'emploi, *Techniques de l'Ingénieur*, tome R2594 :1-26.

[6] Forney, L.J. and Meeks, E.L., Ma, J., Fralick, G.C. 1993. Measurement of frequency response in short thermocouple wires, *Rev. Sci. Intrum.* 64 (5) :1280-1286.

[7] Yule, A.J. and Taylor, D.S., Chigier, N.A. 1978. On-line digital compensation and processing of thermocouples signals for temperature measurements in turbulent flames, *AIAA 16th Aerospace Sciences Meeting*, 78–80.
[8] Lenz, W. and Günther, R. 1980. Measurement of fluctuating temperature in a free-jet diffusion flame, *Comb. and Flame* 37:63-70.
[9] Lockwood, F.C. and Moneib, H.A. 1980. Fluctuating temperature measurements in a heated round free jet, *Comb. Sci. and Technology* 22:63-81.
[10] Voisin, P. and Thiery, L., Brom, G. 1999. Exploration of the atmospheric lower layer thermal turbulences by means of microthermocouples, *E.P.J. App. Phys.* 7(2):177-187
[11] Pitts, W.M. and Braun, E.B., Peacock, R.D., Mitler H.E. et al. 1998. Temperature uncertainties for bare-bead and aspirated thermocouple measurements in fire environnements, in Proceedings of the 14th Meeting of the United States Japan conference on Development of Natural Resources (UJNR) Panel on Fire Research and Safety, May, Japan.
[12] Blevins, L.G. Pitts, W.M. 1999. Modeling of bare and aspirated thermocouples in compartment fires, *Fire Safety J.* 33(4):239-259.
[13] Santoni, P-A. and Marcelli, T., Leoni, E. 2002. Measurement of fluctuating temperatures in a continuous flame spreading across a fuel bed using a double thermocouple probe. *Combustion and Flame* 131(1-2):47-58.
[14] Rakopoulos, C. D. and Rakopoulos, D.C., Mavropoulos, G.C., Giakoumis, E.G. 2004. Experimental and theoretical study of the short term response temperature transients in the cylinder walls of a diesel engine at various operating conditions, *Appl. Therm. Eng.* 24(5-6):679-702.
[15] Bardon, J.P. and Raynaud, M., Scudeller, Y. 1995. Mesures par contact des températures de surface, *Rev. Gén. Therm.* 34(HS95):15-35.
[16] Paranthoen, L. and Lecordier, J.C. 1996. Mesures de température dans les écoulements turbulents, *Rev. Gén. Therm.* 35 :283-308.
[17] Olivari, D. and M. Carbonaro. 1994. *Hot wire measurements. Measurements techniques in fluid dynamics. an introduction*, Von Karman Institut for Fluid Dynamics, *Annual Lecture Series*, vol. 1:183-218.
[18] Million, F., Parenthoën, P., Trinite, M. 1978. Influence des échanges thermiques entre le capteur et ses supports sur la mesure des fluctuations de température dans un écoulement turbulent, *Int. J. Heat Mass Transfer* 21:1-6.
[19] Bradley, D. and Mathews, K. 1968. Measurement of high gas temperature with fine wire thermocouple. J. Mech. Engn. Sci. 10(4):299-305.
[20] Collis, D.C. and Williams, M.J. 1959. Two dimensional convection from heated wires at low Reynolds numbers. *J. of Fluid Mech.* 6:357-384.
[21] Knudsen, J.G. and D.L. Katz. 1958. *Fluid dynamics and heat transfer*, Mc Graw-Hill Book Co., New-York.
[22] Van der Hegg Zijnen, B.G. 1956. Modified correlation formulae for the heat transfer by natural and by forced convection from horizontal cylinders. *Appl. Sci. Res.* A(6):129-140.
[23] Mac Adams, W.H. 1956. *Heat transmission*, Mc Graw-Hill Book Co., New-York.
[24] Eckert, E. R. and Soehngen, E. 1952. Distribution of heat transfer coefficients around circular cylinders, Reynolds numbers from 20 to 500, *Trans. ASME, J. Heat Transfer*, 74:343-347.
[25] Scadron, M.D. and Warshawski, I. 1952. Experimental determination of time constants and Nusselt numbers for bare-wire thermocouples in high velocity air streams and analytic approximation of conduction and radiation errors, NACA, T.N., 2599.
[26] Tarnopolski, M.and Seginer, I. 1999. Leaf temperature error from heat conduction along the wires, *Agr. For. Meteo.*, 93(3):185-190.
[27] Bailly, Y. 1998. Analyse expérimentale des champs acoustiques par méthodes optiques et microcapteurs de température et de pression, Ph. D. diss, University of Franche-Comté, France.

[28] Fralick, G.C. and Forney, L.J. 1993. Frequency response of a supported thermocouple wire: effects of axial conduction, *Rev. Sci. Intrum.* 64(11):3236-3244.
[29] Sbaibi, H. 1987. Modélisation et étude expérimentale de capteurs thermiques, Ph. D. diss, University of Rouen, France.
[30] Singh, B.S. and Dybbs, A. 1976. Error in temperature measurements due to conduction along the sensor leads, *J. Heat Transfer* 491:491-495.
[31] Kramers, H. 1946. Heat transfer from spheres to flowing media, *Physica* 12:61-80.
[32] King, L.V. 1914. On the Convection of Heat from Small Cylinders in a Stream of Fluid, *Phil. Trans. of Roy. Soc. (London)*, Ser. A., 214(14):373-432.
[33] Hilaire, C. and Filtopoulos, E., Trinite, M. 1991. Mesure de température dans les flammes turbulentes. Développement du traitement numérique du signal d'un couple thermoélectrique. *Rev. Gén. Therm.* 354/355 :367-374.
[34] Castellini, P. and Rossi, G.L. 1996. Dynamic characterization of temperature sensors by laser axcitation, *Rev. Sci. Instrum.* 67(7):2595-2601.
[35] Hostache, G. and Prenel J.P., Porcar, R. 1986. Couples thermoélectriques à définition spatiotemporelle fine. Réalisation. Réponse impulsionnelle de microjonctions cylindriques. *Rev. Gén. Therm.* 299 :539-543.
[36] Bardon J P 2001 « *Mesure de température et de flux de chaleur par des méthodes par contact*», Lecture c2b, Ecole d'Hiver METTI , Odeillo, 25-30 jan 1999, Vol.1, (Perpignan: Presse Univ).
[37] Cassagne B, Bardon J P and Beck J V 1986 «*Theoretical and experimental analysis of two surface thermocouple*», Int. Heat Transfer Conf., San Fransisco.
[38] Cassagne B, Kirsch G and Bardon JP 1980 *Int. J. Heat Transfer* **23** 1207-1217
[39] Ravaltera G, Cornet M, Duthoit B and Thery P 1982 *Revue Phys. Appl*.. **17** 4 177-185
[40] Beck J V, Blackwell B and StClair C.A., 1985 Inverse heat conduction (New York: Wiley)
[41] Alifanov O.M 1990 Inverse heat transfer problems (Springer)
[42] Ozisik N 1993 Heat conduction 2d ed.. (New York: Wiley)
[43] Jarny Y, Ozisik M.N and Bardon JP 1991 Int J Heat Mass Transfer 34,11, 2911-2919
[44] Bourouga B., Goizet V and Bardon J P 2000 Int. J. Therm. Sci. 39 96-109

Lecture 3. Basics for linear estimation, the white box case

F. Rigollet[1], D. Maillet[2]

[1] Aix Marseille Université, IUSTI UMR CNRS 7343,
5 rue Enrico Fermi,13453 Marseille cedex 13, France
E-mail: fabrice.rigollet@ univ-amu.fr
[2] Université de Lorraine, LEMTA UMR CNRS 7563, 2 av. de la Forêt de Haye
54504 Vandoeuvre cedex, France
E-mail: denis.maillet@univ-lorraine.fr

Abstract. We present and illustrate the roadmap for a parameter estimation problem from experimental data using a forward model that links the data and the parameters. We focus on the case when the structure of the model is known ('white box case') and linear. The Ordinary Least Square case is considered to introduce all the useful tools (sensitivity coefficients, conditioning of sensitivity matrix, etc). We focus then on optimal ways to implement the best estimation through the study of the sensitivity matrix and other matrices depending on it. The propagation of bias on blocked parameter during the estimation of desired parameters is also studied. The design of optimal experiment (tuning of experiment control parameters) is also presented, based on criterion built on sensitivity matrix and covariance matrix of estimated parameters.

1. Introduction

In experimental heat transfer, we often have to process observed data (measured quantities : temperatures, electrical tensions, radiative flux etc) collected for different values of experimental control variables (time, space, frequencies, wavelength, etc) in order to deduce other quantities of interest (called here *parameters* : thermal diffusivities/effusivities, surface heat flux, internal heat source, emissivities or even temperatures). In parallel of the true experiment that provide the measured quantities (red branch of ***Figure 1***), we suppose we are able to build a set of relations (physical equations, calibration functions) that link the parameters (inputs or *causes*) to the quantities that will be measured on the ground (output or *effects*). This model of the experiment (green branch of ***Figure 1***) is then able to predict the effects for given causes: as it works in the same 'natural' sense causes→effects than the experimental process, it is called the *forward* model. The problem we want to solve is called *inverse measurement problem* because we feed it with the measured quantities (observed data) and we expect it to provide the parameters of interest present in the model: it works in the sense effects→causes, *inverse* of the natural sense. The way it works is based on the comparison between the data measured with the true experiment and the data predicted by the forward model: the optimal *estimated* parameters will be those for which the data predicted by the model are closest to the measured ones (minimization of cost function on

the blue branch of ***Figure 1***). The question that is finally posed deals with the confidence that can be associated to these estimated parameters: how far are they from the (unknown) exact parameters, is their estimated values highly or poorly affected by the random part (noise) of measurement? The last question we can ask ourselves is in fact a question that comes upstream, before the experiment is carried out: can we design the best experiment, that is the experiment that will enable the estimation of parameter with the best confidence (magenta branch of ***Figure 1***, that uses the forward model to answer that question).

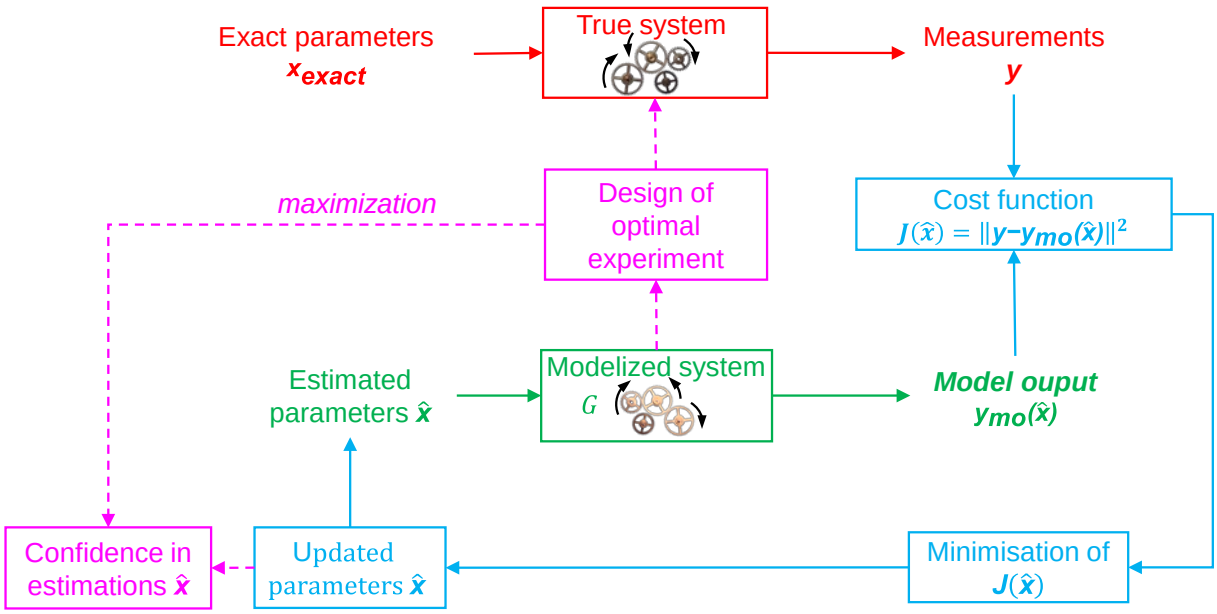


Figure 1 : diagram summarizing the three components of an inverse measurement problem : carrying out an experiment on a true system (red branch), developing a model to simulate the experiment (green branch), and minimizing a cost between the two (blue branch) in order to estimate some parameters of the model, together with their confidence level. A fourth branch is also presented consisting in the design of optimal experiment, whose objective is to design the experiment that will provide the parameters with the best confidence.

We will develop in the following sections all these concepts with the two strong assumptions:

- The *white box*[2] assumption: we suppose we have the right forward model (we have understood the physical phenomenon and the way the measurements are obtained), we do not search its structure, but only its input parameters.
- The *linear* assumption: the model output is a linear function of its parameters see Lecture 7 for non-linear case)

Other assumptions (on noise measurements) or choices (on cost function, on experimental control variables) will be detailed when needed. This rodmap for inverse parameter estimation in the linear case will be illustrated in different situations.

[2] In opposition to the *black box* assumption for which we do not know the structure of the relations between the input parameters and the output of the forward model

2. The roadmap for solving a linear parameter estimation problem: the Ordinary Least Square case

2.1. Generate data: run the true experiment

Let us suppose we have realised an experiment that provides m measurements $y_i = y(t_i)$ for i =1, …, m at m discrete values of time t (the 'independent' variable). These measurements are the components of the vector ($m \times 1$) of experimental measurements $\boldsymbol{y} = [y_1 \ldots y_i \ldots y_m]^t$. Times of measurements are regularly spaced between $t_{\min}$ and $t_{\max}$ and are the components of the time vector ($m \times 1$) $\boldsymbol{t} = [t_{\min} \ldots t_i \ldots t_{\max}]^t$ with $t_i = t_{min} + (i-1)\,\mathrm{d}t$, i=1, …, m. Let ε_i be the (unknown) error associated to the measurement y_i (i=1, …, m), then the measurement errors vector ($m \times 1$) is $\boldsymbol{\varepsilon} = [\varepsilon_1 \ldots \varepsilon_i \ldots \varepsilon_m]^t$. Some assumptions have to be done on these measurement errors. They are detailed in **Table 1**.

Number	Assumption	Explanation
1	Additive errors	$\boldsymbol{y} = \boldsymbol{y}_{perfect} + \boldsymbol{\varepsilon}$
2	Unbiased model	$\boldsymbol{y}_{perfect} = y_{mo}(x^{exact})$
3	Zero mean errors	$\mathrm{E}[\boldsymbol{\varepsilon}] = 0$
4	Constant variance	$\mathrm{var}[\varepsilon] = \sigma_\varepsilon^2$
5	Uncorrelated errors	$cov\ [\varepsilon_i,\ \varepsilon_j] = 0$ for $i \neq j$
6	Normal probability distribution	
7	Known parameters of the probability density distribution of errors	
8	No error in the S_{ij}	$\boldsymbol{S}$ is not a random matrix
9	No prior information regarding the parameters	

Table 1 : Statistical assumptions regarding the measurement errors

The first assumption on measurement errors is that they are purely additive :

$$\boldsymbol{y} = \boldsymbol{y}_{perfect} + \boldsymbol{\varepsilon} \qquad \textbf{(3.1)}$$

Here $\boldsymbol{y}_{perfect}$ represents the vector ($m \times 1$) of (unknown) errorless measurements, which corresponds to the output of a model that is assumed to be perfect[3]. Moreover, measurement errors are assumed to be the realizations of a random variable with any distribution but with a zero mean, that is $\mathrm{E}[\boldsymbol{\varepsilon}] = 0$ (unbiased errors), E[.] being the expected value operator (representing the mean of a large number of realizations of the random

[3] The objective of 'direct' modelisation is to give the best approximation of $\boldsymbol{y}_{perfect}$

variable). On its main diagonal, the covariance matrix ($m \times m$) $\boldsymbol{\psi} = \mathrm{cov}(\boldsymbol{\varepsilon}) = \mathrm{E}\left[(\boldsymbol{\varepsilon} - \mathrm{E}[\boldsymbol{\varepsilon}])\,(\boldsymbol{\varepsilon} - \mathrm{E}[\boldsymbol{\varepsilon}])^{\mathrm{t}}\right] = \mathrm{E}\left[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^{t}\right]$ of measurements errors contains the variance σ_ε^2 of each measurement that is supposed constant for each time t_i, i = 1, …, m. This variance may or may not be known. Finally, measurement errors are assumed uncorrelated (error at time t_i is independent of error at time t_j ($\mathrm{E}[\varepsilon_i\,\varepsilon_j] = 0$ for $i \neq j$) and consequently $\boldsymbol{\psi}$ is a diagonal matrix:

$$\boldsymbol{\psi} = \mathrm{cov}\,(\boldsymbol{\varepsilon}) = \mathrm{diag}\,\left(\sigma_\varepsilon^2, \cdots, \sigma_\varepsilon^2, \cdots, \sigma_\varepsilon^2\right) = \sigma_\varepsilon^2\,\boldsymbol{I} \qquad \textbf{(3.2)}$$

These data (3.1) can come from a real experiment or can have been numerically created (for testing the parameter estimation method), using a mathematical model and adding a numerical random noise verifying the preceding assumptions (and called the *Gaussian* independent and identically distributed (i. i. d.) noise)). Now the model and its parameters will be presented.

2.2. Build a model of the measured signal, define the parameters and first contact with the sensitivities

The objective of such a model is to give a mathematical expression $y_{mo}(t, \boldsymbol{x}) = \eta\,(t, \boldsymbol{x})$, noted $y_{perfect}\,(t)$ of the perfect measurements mentioned above. This model is a function of the independent variable (time) and of n parameters $\boldsymbol{x} = [x_1 \ldots x_n]^t$ composing the parameters vector ($n \times 1$) noted. The model vector ($m \times 1$) is then given by $\boldsymbol{y_{mo}}(\boldsymbol{t}, \boldsymbol{x}) = \left[y_{mo,1}(t_1, \boldsymbol{x}) \ldots y_{mo,i}(t_i, \boldsymbol{x}) \ldots y_{mo,m}(t_m, \boldsymbol{x})\right]^t$, where $\boldsymbol{t} = [t_1 \ldots t_i \ldots t_m]^t$ is a column vector composed of the m times of measurements t_i. For this example, we choose to analyse the classical two parameters estimation problem consisting in estimating simultaneously the slope and the intercept of a straight line; then the model is given, in a scalar writing, by:

$$y_{mo}(t, \boldsymbol{x}) = x_1\,t + x_2 \qquad \textbf{(3.3)}$$

The model is linear with respect to its two parameters x_1 and x_2 because:

$$\begin{aligned} y_{mo}(t, a\boldsymbol{x} + b\boldsymbol{x}') &= (a\,x_1 + b\,x_1')\,t + a\,x_2 + b\,x_2' = a(x_1 t + x_2) + b\,(x_1'\,t + x_2') \\ y_{mo}(t, a\boldsymbol{x} + b\boldsymbol{x}') &= a\;y_{mo}(t, \boldsymbol{x}) + b\;y_{mo}(t, \boldsymbol{x}') \end{aligned} \qquad \textbf{(3.4)}$$

Important remark: the following model:

$$y_{mo}(t, \boldsymbol{x}) = x_1\sqrt{t} + x_2\,\mathrm{erf}\,(t) \qquad \textbf{(3.5)}$$

is also linear with respect to its two parameters x_1 and x_2, even if its time behavior is not linear. On the contrary, the following model:

$$y_{mo}(t, \boldsymbol{x}) = x_1\sqrt{t} + \exp(-x_2\,t) \qquad \textbf{(3.6)}$$

is linear with respect to X_1 but nonlinear with respect to X_2 and is consequently nonlinear with respect to $\boldsymbol{x}$.

Writing the *m* model values (3.3) for the *m* time values $t_1 \ldots t_m$, the *m* resulting equations can be written in a matrix way as follows:

$$\begin{bmatrix} y_{mo,1} \\ \vdots \\ y_{mo,i} \\ \vdots \\ y_{mo,m} \end{bmatrix} = \begin{bmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_i & 1 \\ \vdots & \vdots \\ t_m & 1 \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \textbf{(3.7)}$$

or, in a more compact form:

$$\boldsymbol{y_{mo}} = \boldsymbol{S}\ \boldsymbol{x} \quad \text{whose i}^{\text{th}} \text{ component (i=1 to m) is } y_{mo,i}(t_i,\boldsymbol{x}) = \sum_{k=1}^{n} S_k(t_i) x_k \qquad \textbf{(3.8)}$$

The matrix $\boldsymbol{S}$ ($m \times n$) is called the sensitivity (or Jacobian) matrix. Column *k* contains the *m* times values of the sensitivity coefficient of the model with respect to the parameter X_k , given by :

$$S_k(t,\boldsymbol{x}) = \left. \frac{\partial y_{mo}(t,\boldsymbol{x})}{\partial x_k} \right|_{t,x_j \text{ for } j \neq k} \quad , \quad k=1, \ldots, n \qquad \textbf{(3.9)}$$

Equation (3.8) is only valid for a linear model. However, the sensitivity coefficient (3.9) can be defined for the discrete time values $t = t_i$ (i =1, …, *m)* to form a sensitivity matrix $\boldsymbol{S}$ defined for any linear or nonlinear model as:

$$\mathbf{S}(\boldsymbol{x}) = \left(\boldsymbol{\nabla_x} \boldsymbol{y}_{\boldsymbol{mo}}^{t}\right)^{t} \qquad \textbf{(3.10a)}$$

or, more simply, in a symbolic way

$$\boldsymbol{S}\ (\boldsymbol{x}) = \frac{\mathrm{d}\boldsymbol{y}_{mo}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} \qquad \textbf{(3.10b)}$$

Let us note here that the nabla operator $\boldsymbol{\nabla_x}$, of dimensions n x 1 , can be applied either to a scalar or to a row vector. So,

$$\boldsymbol{\nabla}_{\boldsymbol{x}} z = \begin{bmatrix} \dfrac{\partial z}{\partial x_1} \\ \dfrac{\partial z}{\partial x_2} \\ \vdots \\ \dfrac{\partial z}{\partial x_n} \end{bmatrix} \quad \text{if z is a scalar} \qquad \textbf{(3.10c)}$$

$$\boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{z} = \begin{bmatrix} \dfrac{\partial z_1}{\partial x_1} & \dfrac{\partial z_2}{\partial x_1} & \cdots & \dfrac{\partial z_m}{\partial x_1} \\ \dfrac{\partial z_1}{\partial x_2} & \dfrac{\partial z_2}{\partial x_2} & \cdots & \dfrac{\partial z_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial z_1}{\partial x_n} & \dfrac{\partial z_2}{\partial x_n} & \cdots & \dfrac{\partial z_m}{\partial x_n} \end{bmatrix} \quad \text{if it is a row vector noted } \boldsymbol{z} \text{ of size } 1 \, x \, m \qquad \textbf{(3.10d)}$$

Let us note here that, in terms of dimensions, the left product of the operator ∇_x with another quantity (scalar or matrix), respects the same rule as a normal column vector. For example, the dimensions of $\nabla_x \boldsymbol{z}$ in equation (3.10d) is (n x 1) by (1 x m), that is (n x m).

Important remark: if the model is linear with respect to its parameters (as in the cases (3.3) and (3.5)), then the sensitivity coefficients do not depend on the parameters, and the sensitivity matrix does not depend on $\boldsymbol{x}$.

For model (3.3), we have $S_1(t) = t$ and $S_2(t) = 1$ then:

$$\boldsymbol{S} = \begin{bmatrix} S_1(t_1) & S_2(t_1) \\ \vdots & \vdots \\ S_1(t_i) & S_2(t_i) \\ \vdots & \vdots \\ S_1(t_m) & S_2(t_m) \end{bmatrix} = \begin{bmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_i & 1 \\ \vdots & \vdots \\ t_m & 1 \end{bmatrix} \qquad \textbf{(3.11)}$$

A sensitivity coefficient is a measure of the "influence" of a given parameter X_k on the response of the model $\boldsymbol{y_{mo}}(\boldsymbol{t}, \boldsymbol{x})$. If all the sensitivity coefficients are of "high" magnitude and "independent", the simultaneous estimation of the parameters composing $\boldsymbol{x}$ will be possible. The meaning of "high" and "independent" will be developed later.

2.3. Choose the objective (or "cost") function

The "white box assumption" considers that the model has the right form (or "right structure", given for example by the resolution of the "right" partial differential equations describing the

"right" physical phenomena) then if it is calculated with the right values of parameters $\boldsymbol{x}^{exact}$, we have $\boldsymbol{y}_{mo}(\boldsymbol{x}^{exact}) = \boldsymbol{y}_{perfect}$ and Eq. (3.1) becomes

$$\boldsymbol{y} = \boldsymbol{y}_{mo}(\boldsymbol{x}^{exact}) + \boldsymbol{\varepsilon} \qquad \textbf{(3.12)}$$

Since the m measurement errors composing $\boldsymbol{\varepsilon}$ are not known, the problem of finding the values of the n components of $\boldsymbol{x}^{exact}$ given m measurements verifying Eq. (3.12) is underdetermined (m equations with $n + m$ unknowns: n parameters x_k (k=1, …, n) and m noise values ε_i (i = 1, …, m)). The problem consists in using the m measurements for estimating the n unknown parameters, with $m \geq n$. Then the new problem to solve is a minimization problem. For a given value $\boldsymbol{x}$ of the parameter vector, a residual vector $\boldsymbol{r}$ ($m \times 1$) is built in order to calculate the difference between measurement vector $\boldsymbol{y}$ ($m \times 1$) and the corresponding model output $\boldsymbol{y}_{mo}(\boldsymbol{x})$ (m x 1), each component of $\boldsymbol{r}$ being associated with one of the m instants of time where a measurement is available.

$$\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{x}) = \left[y_1 - y_{mo,1}(t_1, \boldsymbol{x}) \quad \ldots \quad y_i - y_{mo,i}(t_i, \boldsymbol{x}) \quad \ldots \quad y_m - y_{mo,m}(t_m, \boldsymbol{x}) \right]^t \qquad \textbf{(3.13)}$$

This present definition of the residual vector $\boldsymbol{r}(\boldsymbol{x})$ is an extension of the concept of residual vector which is usually defined as $\boldsymbol{r}(\hat{\boldsymbol{x}})$, where $\hat{\boldsymbol{x}}$ corresponds to the minimum of $\boldsymbol{r}(\boldsymbol{x})$, see Eq. (3.18) further on.

Then the norm of this residual vector $\left\| \boldsymbol{r}(\boldsymbol{x}) \right\|$ is calculated, it is a scalar value that will be minimized with respect to the different components of parameter $\boldsymbol{x}$ in order to estimate an 'optimal' value for it. One has to choose the way of computing the norm of the residuals $\left\| \boldsymbol{r}(\boldsymbol{x}) \right\|$. Without any a priori information about the values of the parameters and given the above assumptions for measurements errors, the chosen norm is the Euclidian norm (or L_2 norm) given by:

$$\left\| \boldsymbol{r}(\boldsymbol{x}) \right\| = \left(\sum_{i=1}^{m} r_i^2(\boldsymbol{x}) \right)^{1/2} \qquad \textbf{(3. 14)}$$

In fact, the objective function that will be minimized is the square of that Euclidian norm, it is called the 'Ordinary Least Squares' (OLS) objective (or cost) function[4] :

$$J_{OLS}(\boldsymbol{x}) = \left\| \boldsymbol{r}(\boldsymbol{x}) \right\|^2 = \left\| \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{x}) \right\|^2 \qquad \textbf{(3.15)}$$

[4] it is here the most *efficient*, *i.e.* that will provide the estimation with the minimal variance if the noise is of zero mean, independent and identically distributed

In the particular case of a linear model, $\boldsymbol{y}_{mo}(\boldsymbol{x}) = \boldsymbol{S}\,\boldsymbol{x}$ and this OLS sum becomes:

$$J_{OLS}(\boldsymbol{x}) = \sum_{i=1}^{m} r_i^2(\boldsymbol{x}) = \sum_{i=1}^{m} \left(y_i - y_{mo}(t_i, \boldsymbol{x}) \right)^2 = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} S_j(t_i)\, x_i \right)^2 \qquad \textbf{(3.16)}$$

With a matrix writing, (3.16) is equivalent to :

$$J_{OLS}(\boldsymbol{x}) = \left[\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{x})\right]^t \left[\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{x})\right] \qquad \textbf{(3.17)}$$

The solution of this minimization will be called $\hat{\boldsymbol{x}}_{OLS}$ here :

$$\hat{\boldsymbol{x}}_{OLS} = \arg\left(\min\left(J_{OLS}(\boldsymbol{x})\right)\right) \qquad \textbf{(3.18)}$$

The hat ($\wedge$) superscript designates an estimator of the corresponding quantity, that is a random variable derived here from the random vector variable $\boldsymbol{\varepsilon}$ (the measurement noise) and the subscript 'OLS' designates the specific minimized norm used here, the *Ordinary Least Squares* sum J_{OLS} defined in Eq. (3.14). If the model is linear, this OLS estimator does not require the use of any iterative algorithm and is given in a simple explicit form that will be presented later.

To summarize, the original question was:

> "what are the exact values $\boldsymbol{x}^{exact}$ of parameter vector $\boldsymbol{x}$ for the model $\boldsymbol{y}_{mo}(\boldsymbol{x})$ when *m* corresponding noisy measurements $\boldsymbol{y} = \boldsymbol{y}_{mo}(\boldsymbol{x}^{exact}) + \boldsymbol{\varepsilon}$ are available?"

The answer is:

> "one possible approximation of $\boldsymbol{x}^{exact}$ is the estimator $\hat{\boldsymbol{x}}_{OLS}$, which minimizes the Ordinary Least Squares 'objective' function (sometimes also called 'criterion') $J_{OLS}(\boldsymbol{x})$ defined as the sum of the squares of the differences between the *m* model output and the corresponding measurements".

Or, in simpler words:

> "the natural numerical approximation of the parameters present in $\boldsymbol{x}^{exact}$ is the one that enables the model to be the closest to the whole set of measurements. This Ordinary Least Squares method was first found by Carl Friedrich Gauss in 1795 and later published by Adrien-Marie Legendre (1805)".

The natural question that arises next is: "how far is this $\hat{\boldsymbol{x}}_{OLS}$ estimation from the exact value $\boldsymbol{x}^{exact}$ and what can be done to reduce their difference?" These questions will be discussed

now within the linear assumption where an explicit expression for $\hat{\boldsymbol{x}}_{OLS}$ will be given. Readers interested by non-linear estimation can refer to lecture 7 of this series.

2.4. Solve the parameter estimation problem: minimize the objective function

The OLS estimator $\hat{\boldsymbol{x}}_{OLS}$ is defined as the value of parameter vector $\boldsymbol{x}$ that minimizes the scalar objective function $J_{OLS}(\boldsymbol{x})$. Then, for $\boldsymbol{x}=\hat{\boldsymbol{x}}_{OLS}$, the gradient of $J_{OLS}(\boldsymbol{x})$ must be zero :

$$\boldsymbol{\nabla}_x J_{OLS}(\boldsymbol{x}) = 0 \tag{3.19}$$

$$\boldsymbol{\nabla}_x J_{OLS}(\boldsymbol{x}) = 2\left(\boldsymbol{\nabla}_x(\boldsymbol{y}\text{-}\boldsymbol{y}_{mo}(\boldsymbol{x}))^t\right)(\boldsymbol{y}\text{-}\boldsymbol{y}_{mo}(\boldsymbol{x})) = 0 \tag{3.20}$$

This equation stems from the following property of the nabla operator $\boldsymbol{\nabla}_x$, applied to a scalar product of vectors, see (Beck and Arnold, 1977, page 221) in the reference list[5] :

$$\boldsymbol{\nabla}_x(\boldsymbol{z}^t\boldsymbol{z}) = 2(\boldsymbol{\nabla}_x\boldsymbol{z}^t)\boldsymbol{z} \tag{3.21}$$

where $\boldsymbol{z}$ is a column-vector of size (m x 1) then $\boldsymbol{z}^t$ is a line-vector of size (1 x m). Reminding the linear model expression (3.8) $\boldsymbol{y}_{mo} = \boldsymbol{S}\,\boldsymbol{x}$, the definition of the sensitivity matrix (3.10a) and its transpose $\boldsymbol{S}^t = \boldsymbol{\nabla}_x \boldsymbol{y}_{mo}^t$ and knowing that $\boldsymbol{\nabla}_x \boldsymbol{y}^t = \boldsymbol{0}$ because measurements $\boldsymbol{y}$ do not depend on parameters $\boldsymbol{x}$, we can write

$$\boldsymbol{\nabla}_x(\boldsymbol{y}\text{-}\boldsymbol{y}_{mo}(\boldsymbol{x}))^t = \boldsymbol{\nabla}_x\boldsymbol{y}^t - \boldsymbol{\nabla}_x\boldsymbol{y}_{mo}(\boldsymbol{x})^t = 0 - \boldsymbol{S}^t \tag{3.22}$$

and Eq. (3.20) becomes:

$$\nabla_x J_{OLS}(\boldsymbol{x}) = -2\,\boldsymbol{S}^t\left[\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}\right] = 0 \tag{3.23}$$

Then $\hat{\boldsymbol{x}}_{OLS}$ is solution of:

$$\boxed{\left[\boldsymbol{S}^t\,\boldsymbol{S}\right]\hat{\boldsymbol{x}}_{OLS} = \boldsymbol{S}^t\,\boldsymbol{y}} \tag{3.24}$$

The n equations composing the linear system (3.24) are called the 'normal equations'. The solution is straightforward if the ($n \times n$) matrix $\boldsymbol{S}^t\boldsymbol{S}$ is not singular, it is then possible to compute its inverse and obtain:

[5] it corresponds to the matrix formulation of the derivation rule for a composed function of x : $[f(x)^2]' = 2f'(x)f(x)$

$$\boxed{\hat{\boldsymbol{x}}_{OLS} = \left[\boldsymbol{S}^t\boldsymbol{S}\right]^{-1}\boldsymbol{S}^t\,\boldsymbol{y}} \qquad \textbf{(3.25)}$$

Let us note that it is not necessary to invert matrix $\boldsymbol{S}^t\boldsymbol{S}$ in order to solve the system of normal equations (3.24). Equation (3.25) can be used further on to yield a symbolic explicit expression of the OLS solution.

The ($n \times m$) matrix $\left[\boldsymbol{S}^t\boldsymbol{S}\right]^{-1}\boldsymbol{S}^t$ is called the Moore-Penrose matrix, also named as the pseudo-inverse of $\boldsymbol{S}$[6]. Obviously, a necessary condition for $\boldsymbol{S}^t\boldsymbol{S}$ not to be singular is that the sensitivity coefficients are independent, and have a non-zero norm. This condition also requires that the number of measurements m be equal or greater than the number of parameters n to be estimated.

Eq. (3.24) gives an explicit expression for the ordinary least square *estimator* $\hat{\boldsymbol{x}}_{OLS}$ of $\boldsymbol{x}$ for any linear model $\boldsymbol{y}_{mo}(\boldsymbol{x}) = \boldsymbol{S}\,\boldsymbol{x}$ as a function of measurements $\boldsymbol{y}$ defined in Eq. (3.12). Since $\boldsymbol{y}$ is a random vector (because of noise $\boldsymbol{\varepsilon}$), such is also the case for $\hat{\boldsymbol{x}}_{OLS}$. However, equation (3.24) has also another *statistical* meaning: once measurements are available, a realization of $\boldsymbol{y}$ (that is numerical values for its components) becomes available, and this equation provides the corresponding OLS *estimation* of $\boldsymbol{x}$.

2.5. Evaluate the confidence in estimations (variance and bias of estimator)

2.5.1. First approach with stochastic simulations (Monte Carlo method)

Before computing the statistical properties of the OLS estimator (expected value and covariance matrix), we present a graphical approach that helps to understand the meaning of such properties. This approach is possible in the case when two parameters are estimated because each estimation $\hat{\boldsymbol{x}}_{OLS} = (\hat{x}_{OLS,1}, \hat{x}_{OLS,2})$ can be plotted as a point in a 2D coordinates frame graduated in (x_1, x_2). The idea is then to simulate K=100 experiments with K different realizations of the random noise vector $\boldsymbol{\varepsilon}$ generated by an independently distributed Gaussian process with the same statistical properties (see Table 1) to produce K samples of measurements vectors $\boldsymbol{y}$ according to (3.12). The exact output of the model ($\boldsymbol{y}_{perfect}$) as well as the time of measurements and the standard deviation of the noise used for each simulation is given in *Table 2*. This model structure with this set of associated experimental parameters is called the 'reference case'. **Figure 2** shows one of the simulated experiments (circles) and the corresponding recalculated model output corresponding to the OLS estimation $\hat{\boldsymbol{x}}_{OLS} = (\hat{x}_{OLS,1}, \hat{x}_{OLS,2})$ (red line).

[6] if there is as much measurements m than the number n of parameters to estimate, then $\boldsymbol{S}$ is square, its inverse $\boldsymbol{S}^{-1}$ exists and its pseudo-inverse is equal to $\boldsymbol{S}^{-1}$, then the solution is simply $\boldsymbol{x}_{OLS}=\boldsymbol{S}^{-1}\boldsymbol{y}$. In other words : least squares is a tool for (pseudo)inverting overdetermined problems (more data then more equations than parameters).

x_1^{exact} (K/h)	5
x_2^{exact} (K)	2
Model structure $y_{mo}(t,\boldsymbol{x})$, K	$x_1\ t + x_2$, Eq. (3.3)
Number of measurements m	20
Start of time range t_{min} , s	0.5
Time step dt, s	0.1
Noise standard deviation σ_ε , K	0.5

Table 2 : conditions of the K=100 simulated 'reference' experiments.

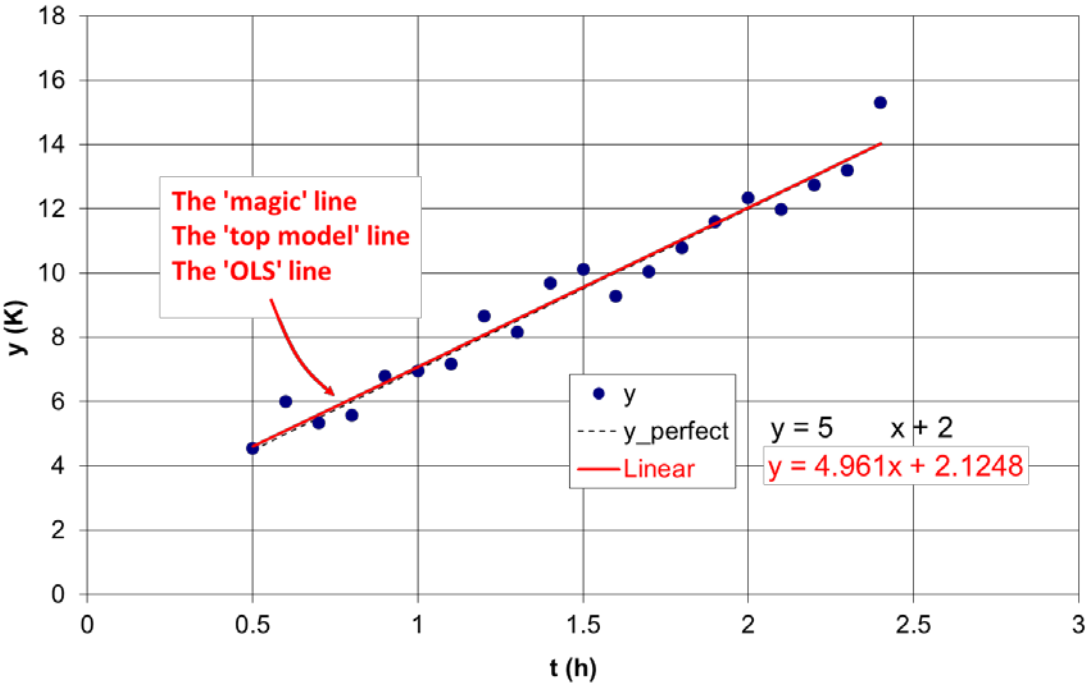


Figure 2 : one of the K=100 experiments of the 'reference case', the corresponding exact model and the corresponding recalculated OLS model output.

The K=100 OLS estimations $\hat{\boldsymbol{x}}_{OLS} = (\hat{x}_{OLS,1}, \hat{x}_{OLS,2})$ are then plotted in a scatter graph graduated in (x_1, x_2) in ***Figure 3***. Because of a different random realization of noise for each of the 100 experiments, each corresponding OLS estimations $\hat{\boldsymbol{x}}_{OLS} = (\hat{x}_{OLS,1}, \hat{x}_{OLS,2})$ is different, showing immediately the consequence of noise measurement on the dispersion of estimations. In that figure the position (square) of the exact value $\boldsymbol{x}^{exact} = (x_1^{exact} = 5, x_2^{exact} = 2)$ and the position (star) of the mean value of the K estimations $\hat{\boldsymbol{x}}_{mean} = (\text{mean}\ (\hat{x}_{1,OLS}) = 4.994, \text{mean}\ (\hat{x}_{2,OLS}) = 2.019)$ (the center of the scatter) are very close.

Another interesting way of looking at the estimation results is to plot them in a scatter graph with normalized coordinates indicating the distance of each estimation from the center of the scatter in %, see ***Figure 4***:

$$e_{OLS,1,i} = 100\left(\hat{x}_{OLS,1,i} - \hat{x}_{mean,1}\right)/\,\hat{x}_{mean,1} \quad \textbf{(3.26)}$$

$$e_{OLS,2,i} = 100\left(\hat{x}_{OLS,2,i} - \hat{x}_{mean,2}\right)/\,\hat{x}_{mean,2} \quad \textbf{(3.27)}$$

If we consider that $\hat{\boldsymbol{x}}_{mean} \approx \boldsymbol{x}_{exact}$ the quantities (3.26) and (3.27) that are the relative estimation errors in % of x_1^{exact} and of x_2^{exact}. That plot enables to quantify in % the dispersion of the estimations of each parameter around its mean value. This dispersion is what one often wants to minimize.

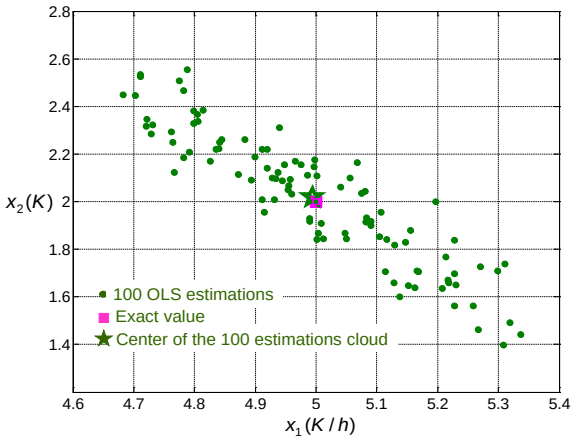


Figure 3 : dispersion of the 100 estimations around their central value (star) that is very close to the exact value (square)

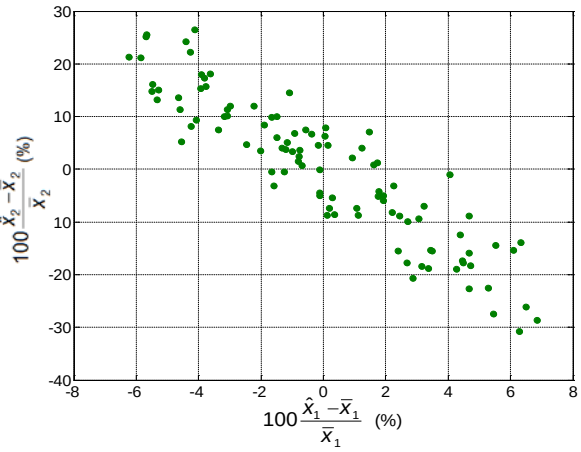


Figure 4* : *relative estimation errors in % centred and scaled using the mean value of the scatter

At this point, after having quantified the central value $\hat{\boldsymbol{x}}_{mean}$ of the K=100 $\hat{\boldsymbol{x}}_{OLS}$ estimations and after having evaluated the dispersion of the majority of estimations around this central value (that indicates the confidence we associate to it), we can sum up the result of the estimation problem in the following way:

"x_1^{exact} is equal to $\hat{x}_{mean,1} = 2.0 \pm 20\%$ and
x_2^{exact} is equal to $\hat{x}_{mean,2} = 4.9 \pm 40\%$"

But actually, we never realize 100 experiments with 100 estimations $\hat{\boldsymbol{x}}_{OLS} = (\hat{x}_{OLS,1}, \hat{x}_{OLS,2})$ in order to calculate the mean value $\hat{\boldsymbol{x}}_{mean} = (\text{mean}(\hat{x}_1), \text{mean}(\hat{x}_2))$. We generally do one single experiment and obtain only one of the 100 points of ***Figure 3*** and ***Figure 4***. We must keep in mind that this point can be one of the points 'far' from the exact value! Whatever the realized

experiment among these 100, what we want to do is to associate a 'confidence region' to the particular estimation $\hat{\boldsymbol{x}}_{OLS} = (\hat{x}_{OLS,1}, \hat{x}_{OLS,2})$ (or 'confidence intervals' for each parameter) that has about the same dimension than the scatter we have just obtained with these 100 simulated experiments. That is the objective of the following section.

2.5.2. Calculation of statistical properties of the OLS estimator

Here we become more general and we consider the case when not all the *n* parameters are estimated but only *r*, the (*n*-*r*) remaining parameters are supposed to be known and they are fixed during the estimation of the *r* unknown parameters. Usually a parameter is set to a supposed known values for two major reasons: i) the model is not sensitive enough to that parameters or ii) the sensitivity of the model to that parameter 'looks like' the sensitivity to another parameter (see Section 3.1.1). Unknown parameters are noted with subscript *r* and known parameters are noted with subscript *c*. We must consider that the fixed parameters have not been fixed to their exact value, and at the end of estimation of the *r* parameters, we have to evaluate the bias made on the estimations because of the error in the (*n* - *r*) parameters that are supposed to be known.

We can split (3.8) into:

$$\boldsymbol{y}_{mo} = \boldsymbol{S}_r \boldsymbol{x}_r + \boldsymbol{S}_c \boldsymbol{x}_c \tag{3.28}$$

The matrix $\boldsymbol{S}_r$ ($n \times r$) is the sensitivity matrix to estimated parameters. It is a part of the "complete" sensitivity matrix $\boldsymbol{S}$, relative to all the parameters (unknown $\boldsymbol{x}_r$ ($r \times 1$) and known $\boldsymbol{x}_c$ ((n-r) × 1)):

$$\boldsymbol{S} = [\boldsymbol{S}_r \vdots \boldsymbol{S}_c] = \left[\begin{bmatrix} S_1(t_1) & \dots & S_r(t_1) \\ \vdots & \dots & \vdots \\ S_1(t_m) & \dots & S_r(t_m) \end{bmatrix} \vdots \begin{bmatrix} S_{r+1}(t_1) & \dots & S_{n-r}(t_1) \\ \vdots & \dots & \vdots \\ S_{r+1}(t_m) & \dots & S_{n-r}(t_m) \end{bmatrix}\right] \tag{3.29}$$

The matrix $\boldsymbol{S}_c$ (n × r) is the sensitivity matrix to estimated parameters. It is a part of the "complete" sensitivity matrix $\boldsymbol{S}$, relative to all the parameters (estimated $\boldsymbol{x}_r$ ($r \times 1$) and fixed $\boldsymbol{x}_c$ ((n-r) × 1)):

The OLS solution (3.25) becomes:

$$\hat{\boldsymbol{x}}_{OLS} = \left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \boldsymbol{S}_r^t (\boldsymbol{y} - \boldsymbol{S}_c \boldsymbol{x}_c) \tag{3.30}$$

Let $\hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c)$ be the estimated parameters for a value of fixed parameters $\tilde{\boldsymbol{x}}_c$ different from their exact value $\boldsymbol{x}_c^{exact}$. Let $\boldsymbol{e}_r$ be the vector (r × 1) of the estimation error (the difference between estimated $\hat{\boldsymbol{x}}_r$ and exact $\boldsymbol{x}_r^{exact}$ values of $\boldsymbol{x}_r$) and let $\boldsymbol{e}_c$ be the deterministic error (the bias) for the fixed values of the parameters that are supposed to be known:

$$\boldsymbol{e}_r = \hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c) - \boldsymbol{x}_r^{exact} \quad \textbf{(3.31)}$$

$$\boldsymbol{e}_c = \tilde{\boldsymbol{x}}_c - \boldsymbol{x}_c^{exact} \quad \textbf{(3.32)}$$

One can write, with $\boldsymbol{A}_r = \left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \boldsymbol{S}_r^t$ the Moore-Penrose matrix:

$$\hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c) = \boldsymbol{A}_r(\boldsymbol{y} - \boldsymbol{S}_c \tilde{\boldsymbol{x}}_c) \quad \textbf{(3.33)}$$

Eq. (3.12) can be developed:

$$\boldsymbol{y} = \boldsymbol{y}_{mo}(\boldsymbol{x}^{exact}) + \boldsymbol{\varepsilon} = \boldsymbol{S}_r \boldsymbol{x}_r^{exact} + \boldsymbol{S}_c \boldsymbol{x}_c^{exact} + \boldsymbol{\varepsilon} \quad \textbf{(3.34)}$$

Combining Eq. (3.34) and (3.33), the estimation error (3.31) may then be approximated by:

$$\boldsymbol{e}_r = \hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c) - \boldsymbol{x}_r^{exact} = \boldsymbol{A}_r \boldsymbol{\varepsilon} - \boldsymbol{A}_r \boldsymbol{S}_c \boldsymbol{e}_c = \boldsymbol{e}_{r1} + \boldsymbol{e}_{r2} \quad \textbf{(3.34)}$$

The first term $\boldsymbol{e}_{r1} = \boldsymbol{A}_r \boldsymbol{\varepsilon}$ is the random contribution to the total error; it represents the error due to measurement errors $\boldsymbol{\varepsilon}$ whose covariance matrix $\boldsymbol{\Psi}$ is given by Eq. (3.2). The second term $\boldsymbol{e}_{r2} = -\boldsymbol{A}_r \boldsymbol{S}_c \boldsymbol{e}_c$ is the non-random (deterministic) contribution to the total error vector due to the deterministic error on the fixed parameters $\boldsymbol{e}_c$. The expected value of $\boldsymbol{e}_{r1}$ is:

$$\mathrm{E}[\boldsymbol{e}_{r1}] = \boldsymbol{A}_r \mathrm{E}[\boldsymbol{\varepsilon}] = 0 \quad \textbf{(3. 35)}$$

meaning that no systematic bias is introduced by the random measurement errors.

Remark: this explains that the mean $\hat{\boldsymbol{x}}_{mean}$ of the 100 scattered estimations in ***Figure 3*** is very close to the exact value $\boldsymbol{x}^{exact}$.

The covariance matrix of $\boldsymbol{e}_{r1}$ is given by:

$$\boldsymbol{C}_1 = \mathrm{cov}(\boldsymbol{e}_{r1}) = \mathrm{E}[\boldsymbol{e}_{r1} \boldsymbol{e}_{r1}^t] = \boldsymbol{A}_r \mathrm{E}[\boldsymbol{\varepsilon}\, \boldsymbol{\varepsilon}^t] \boldsymbol{A}_r^t = \boldsymbol{A}_r \boldsymbol{\psi} \boldsymbol{A}_r^t = \sigma_\varepsilon^2 \left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \quad \textbf{(3.36)}$$

The matrix $\boldsymbol{P}_r = \left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1}$ may thus be seen as the matrix of "amplification" of measurement errors. The expected value of $\boldsymbol{e}_{r2}$ is:

$$\mathrm{E}[\boldsymbol{e}_{r2}] = -\boldsymbol{A}_r \boldsymbol{S}_c \boldsymbol{e}_c = -\left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \boldsymbol{S}_r^t \boldsymbol{S}_c \boldsymbol{e}_c \neq 0 \qquad \textbf{(3.37)}$$

This expected value is different from zero, which means that estimation $\hat{\boldsymbol{x}}_{r,OLS}$ is biased, if the error $\boldsymbol{e}_c$ of the parameters supposed to be known is different from zero itself. This means that in the preceding stochastic simulation if only one part of $\boldsymbol{x}_r$ had been estimated (with a non-zero error on the remaining part $\boldsymbol{x}_c$) the scatter of 100 estimations would not have been centred on $\boldsymbol{x}_r^{exact}$ (see example on **Figure 20**, section 4, page 34). This bias is computed using the corresponding sensitivity coefficients matrix $\boldsymbol{S}_c$. The covariance matrix ((n-*r*)×(*n*-*r*)) of $\boldsymbol{e}_{r2}$ error is $\boldsymbol{C}_2 = \mathrm{cov}(\boldsymbol{e}_{r2}) = 0$ because $\boldsymbol{e}_c$ is not a random error. Finally, the total bias associated to the estimation $\hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c)$ is due to the biased value of $\tilde{\boldsymbol{x}}_c$ and its value is given by:

$$\mathrm{E}[\boldsymbol{e}_r] = \mathrm{E}[\boldsymbol{e}_{r2}] = -\left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \boldsymbol{S}_r^t \boldsymbol{S}_c \boldsymbol{e}_c \qquad \textbf{(3.38)}$$

The matrix $\left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \boldsymbol{S}_r^t \boldsymbol{S}_c = \boldsymbol{P}_r \boldsymbol{S}_r^t \boldsymbol{S}_c$ ($r \times (n\text{-}r)$) may thus be seen as the "amplification" of bias on the fixed parameters $\boldsymbol{e}_c$. For a fixed value of the supposed known parameters $\tilde{\boldsymbol{x}}_c$, the covariance matrix (size ($r \times r$)) of estimation error is:

$$\boldsymbol{C}_r = \mathrm{cov}(\boldsymbol{e}_r) = \mathrm{E}\left[(\boldsymbol{e}_r - \mathrm{E}\left[\boldsymbol{e}_r\right])(\boldsymbol{e}_r - \mathrm{E}\left[\boldsymbol{e}_r\right])^t\right] = \mathrm{cov}(\boldsymbol{e}_{r1}) \qquad \textbf{(3.39)}$$

The coefficients of the covariance matrix of the estimation error are defined by:

$$\boldsymbol{C}_r = \begin{bmatrix} \sigma_1^2 & \mathrm{cov}(e_{r1}, e_{r2}) & \cdots & \mathrm{cov}(e_{r1}, e_{rr}) \\ & \sigma_2^2 & & \mathrm{cov}(e_{r2}, e_{rr}) \\ & & \ddots & \vdots \\ sym & & & \sigma_r^2 \end{bmatrix} = = \sigma_\varepsilon^2 \left[\boldsymbol{S}_r^t \boldsymbol{S}_r\right]^{-1} \qquad \textbf{(3.40)}$$

Its main diagonal elements is composed of the individual variances of the error associated to each component of the estimated vector $\hat{\boldsymbol{x}}_{r,OLS}$ and its other coefficients are the covariance of crossed errors. Equation (3.40) shows that knowledge of the variance of measurement errors σ_ε^2 is needed in order to compute the covariance matrix. If σ_ε^2 is not measured before the experiment, an estimation of it may be obtained at the end of estimation thanks to the final value of the objective function :

$$J_{OLS}(\hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c)) = \sum_{i=1}^{m} r_i^2\,(\hat{\boldsymbol{x}}_{OLS}(\tilde{\boldsymbol{x}}_c)) \qquad \textbf{(3. 41)}$$

In fact, this estimation is based on the fact that, at the end of the estimation, the only difference that subsists between measurements and model (if its structure and its parameters are correct) must be the measurement errors. In fact, exact parameters are not exactly obtained, and the remaining differences between measurements and model are the residuals given by (3.13). If the estimated parameters are not too far from the exact parameters, the residuals should have some statistical properties close to those of measurement errors. That is why a non-biased estimation of σ_ε^2 for the estimation of *r* parameters from the use of *m* measurements is thus given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{J_{OLS}(\hat{\boldsymbol{x}}_{r,OLS}(\tilde{\boldsymbol{x}}_c))}{m-r} \qquad \textbf{(3. 42)}$$

This estimation is only valid for an independent and identically distributed (i.i.d.) noise $\boldsymbol{\varepsilon}$ and if there is no bias in the parameters supposed to be known, that is $\tilde{\boldsymbol{x}}_c = \boldsymbol{x}_c^{exact}$. Let us note that in the case of 'exact matching', where the number of measurements m is equal to the number *r* of parameters that are looked for, both numerator and denominator of equation (3. 42) are equal to zero and, consequently, no information about the noise level can be brought by the calculation of the residuals.

2.5.3. The correlation matrix

The estimation error associated to $\hat{\boldsymbol{x}}_{r,OLS}(i)$ cannot be arbitrarily low independently of the corresponding error in $\hat{\boldsymbol{x}}_{r,OLS}\ (j)$ in the case where $\text{cov}\,(e_{ri}, e_{rj})$: $\hat{x}_{ri}$ and $\hat{x}_{rj}$ are said correlated through the link that exists between their errors. The correlation level between estimations $\hat{x}_{ri}$ and $\hat{x}_{rj}$ is thus measured by the quantity:

$$\rho_{ij} = \frac{\text{cov}(e_{ri}, e_{rj})}{\sigma_i\,\sigma_j} = \frac{C_{r,ij}}{\sqrt{C_{r,ii}C_{r,jj}}} = \frac{P_{r,ij}}{\sqrt{P_{r,ii}P_{r,jj}}} \quad \text{for } i, j = 1,\ldots, r \qquad \textbf{(3.43)}$$

that lies between -1 and 1.

One considers that two estimation errors are highly correlated when $|\rho_{ij}| \geq 0.9$ (Beck et al., 1977). This quantity is independent of the magnitude of measurement errors and corresponds only to the degree of collinearity of the sensitivity coefficients. In the example of ***Figure 4***, $\rho_{12} = -0.99$ indicates that the error in the estimation of the slope (x_1) is highly linked to the error in the estimation of the intercept x_2 and that they will have the opposite sign or variation (if x_1 is over-estimated (resp. under-estimated) then x_2 will be under-

estimated (resp. over-estimated) with a very high level of probability). However, this correlation coefficient does not bring any information about the level of these errors: this is brought by the calculation of their variances, the diagonal coefficients in Eq. (3.36). The high negative coefficient of correlation between the two parameters in our example explains why the scatter of the 100 estimations is contained inside a 'narrow' and 'inclined' ellipse whose main axis has a negative slope in ***Figure 3***.

2.5.4. The confidence region and interval for OLS with Gaussian assumptions

If the noise is Gaussian and i.i.d. the confidence region in the plane $(\hat{x}_1, \hat{x}_2)$ plane in ***Figure 3***, for a given confidence level α is an ellipse (for *n*=2 parameters, see ***Figure 5***). Its equation in $\boldsymbol{\delta x}$ coordinates centered on $\hat{\boldsymbol{x}}_{OLS}$ is:

$$\boldsymbol{\delta x}^t \boldsymbol{S}^t \boldsymbol{S}\, \boldsymbol{\delta x} = \Delta^2$$
$$\Delta^2 = \chi^2_{1-\alpha}(2)\, \sigma_\varepsilon^2 \qquad \textbf{(3. 44)}$$

$\chi^2_{1-\alpha}(2)$ is computed by the function *chi2inv(1-alpha,2)* in MATLAB® (or GNU-Octave) or *LOI.KHIDEUX.INVERSE(1-alpha;2)* in Excel® if we search for the confidence region at a 95% level (α=0.05) for the estimation of 2 parameters.

Typical values for classical confidence intervals are indicated in the Table 3.

1-α	ν=1	ν=2	ν=3	ν=4
68.30%	**1.00**	**2.30**	**3.53**	**4.72**
95.45%	**4.00**	**6.17**	**8.02**	**9.72**
99.73%	**9.00**	**11.83**	**14.16**	**16.25**

Table 3 : **Chi-Square law for given confidence levels (1 - α) and ν degrees of freedom that will be used to compute the size of the ellipsoidal confidence regions. Square root of values in first column gives the classical rules '1σ, 2σ and 3σ'**

σ_ε^2 is the variance of noisy measurements. It is worth noting that the lengths of half axes ρ_1 and ρ_2 in the principal directions of the ellipse are given by:

$$\rho_1 = \Delta / \sqrt{\lambda_1}$$
$$\rho_2 = \Delta / \sqrt{\lambda_2} \qquad \textbf{(3. 45)}$$

λ_1 and λ_2 are the eigenvalues of $\boldsymbol{S}^t\boldsymbol{S}$. The product of these two eigenvalues is equal to the determinant of $\boldsymbol{S}^t\boldsymbol{S}$. Finally, the area of the confidence region inside the ellipse is given by:

$$A = \pi.\rho_1.\rho_2 = \frac{\pi\chi^2_{1-\alpha}(2)\sigma^2_\varepsilon}{\sqrt{\det(\boldsymbol{S}^t\boldsymbol{S})}} = \frac{\pi\chi^2_{1-\alpha}(2)\sigma^2_\varepsilon}{\sqrt{\lambda_1\lambda_2}} \quad \textbf{(3. 46)}$$

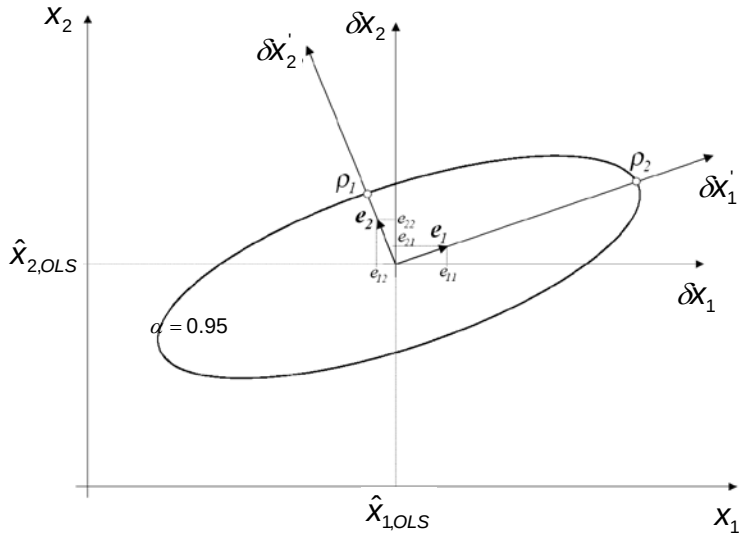


Figure 5 : elliptical confidence region associated to the estimation of two parameters (with Gaussian i.i.d. measurement noise), at a confidence level 1-α=0.95.

So, the product of eigenvalues of $\boldsymbol{S}^t\boldsymbol{S}$ gives information on the area of the confidence region, while the individual eigenvalues give information on the lengths of each principal direction of the ellipse : a 'long' ellipse in a direction corresponds to a low eigenvalue. The experiment that will maximize $det(\boldsymbol{S}^t\boldsymbol{S})=\lambda_1\lambda_2$ in order to minimize the confidence region is called a 'D-optimal' experiment.

In the case of estimation of $r = n$ parameters, the n variances associated to each component of the estimated vector $\hat{\boldsymbol{x}}_{OLS}$ constitute the main diagonal of matrix $\boldsymbol{C}$ (Eq. 3.39). The square root of the ith diagonal component of $\boldsymbol{C}$ is then the standard deviation associated to the estimation $\hat{x}_{i,OLS}$ and can be expressed in %. Then, the half width of confidence interval $CI_i^{1-\alpha}$ at a level of confidence of $100(1-\alpha)\%$, associated to the estimation $\hat{x}_{i,OLS}$ is now given by:

$$CI_i^{1-\alpha} = t_{1-\alpha/2}(m-n)\times\sqrt{C_{ii}}\ ,\ \text{for}\ \ i = 1,\ldots,n \quad \textbf{(3.47)}$$

The quantity $t_{1-\alpha/2}(m-n)$ is the t-statistic for m - n degrees of freedom at the confidence level of $100(1-\alpha)\%$ (function *tinv(1-alpha/2,m-n)* in MALTAB® or *LOI.STUDENT.INVERSE.N(1–alpha/2;m-n)* in Excel®). For example, for m = 20 measurements, if n = 2 parameters are estimated, and if the 95% confidence is wanted, then $\alpha = 0.05$ and $t_{0.975}(20-2) = 2.1$. For a high number of measurements (>200), the t-statistic tends to the Gaussian statistic and we have $t_{0.975} \to 1.96$. Finally, the result of the estimation process of the unknown exact parameter x_i^{exact} can be presented in the following way:

' x_i^{exact} has a 95% chance of being in the interval $\left[\hat{x}_{i,OLS} - CI_i^{0.95} \quad \hat{x}_{i,OLS} + CI_i^{0.95}\right]$,
or: ' $x_i^{exact} = \hat{x}_{i,OLS} \pm CI_i^{0.95}$ with 95% chance'

2.5.5. The residuals analysis

When estimation is achieved, the graphical analysis of residuals given by Eq. (3.13) $\boldsymbol{r}(\hat{\boldsymbol{x}})$ enables to detect some inconsistency of the result. Difference between measurements and model response with optimal parameters must 'look like' measurement noise , or in other words : 'the right model with the right parameters must explain the measurements except its random part'. For a Gaussian noise with standard assumptions, the statistical properties of residuals must be close to the measurement error properties (zero mean and variance (m – n) σ_ε^2). If the residuals are signed, the problem may be due to an error in the statistical assumptions regarding the measurements or in the structure or parameters of the direct model.

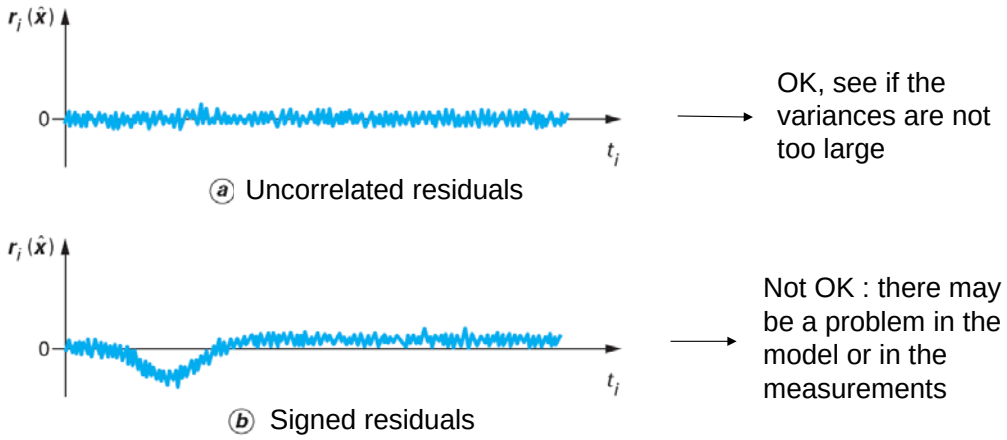


Figure 6 : graphical analysis of residuals at the end of the estimation

3. Indicators for a successful estimation

It has been shown above that matrix $\boldsymbol{S}^t\boldsymbol{S}$, also called the *information matrix*, is fundamental in the process of parameter estimation:

- it has to be invertible (that is non-singular: $\det(\boldsymbol{S}^t\boldsymbol{S}) \neq 0$) in order for **the OLS estimation** to be possible, according to Eq. **(3.25),**
- it also has to be inverted to **compute the covariance matrix** according to Eq. **(3.36)** associated to the OLS estimation. The diagonal terms of this matrix are equal (within the σ_ε^2 factor and in case of an i.i.d. noise) to the variances of each estimation, and the off-diagonal terms enable to compute the correlation matrix. The inverse of $\boldsymbol{S}^t\boldsymbol{S}$ play the role of "noise amplification",

- the eigenvalues of $\boldsymbol{S}^t\boldsymbol{S}$, in the case of a Gaussian i.i.d. noise, enable the **calculation of the lengths of the half principal axes of the elliptical confidence region,**
- the determinant of $\boldsymbol{S}^t\boldsymbol{S}$ enables the **calculation of the area of the elliptical confidence region.**

The difficulty is clear : $\boldsymbol{S}^t\boldsymbol{S}$ has to be non-singular to be inverted and $\boldsymbol{S}^t\boldsymbol{S}$ has to be not 'quasi-singular' in order to limit the noise amplification. This notion of non-singular character of the information matrix $\boldsymbol{S}^t\boldsymbol{S}$ makes sense only if all the parameters x_j have the same physical units. Otherwise, one should study matrix $\boldsymbol{S}^{*t}\boldsymbol{S}^*$ where $\boldsymbol{S}^*$ is the reduced (sometimes called 'scaled') sensitivity matrix, see section Sections 3.1 and 3.3**.**

We then have to find some indicators to evaluate the singularity and the quasi-singularity of $\boldsymbol{S}^{*t}\boldsymbol{S}^*$. The first indication can be simply graphical. Indeed, the singularity would happen if a sensitivity coefficient $S^*_i(t)$ was purely proportional to another $S^*_j(t)$; in that case the rank of $\boldsymbol{S}^{*t}\boldsymbol{S}^*$ is lower than n, and its determinant is zero. More difficult is to find a linear combination of more than two sensitivity coefficients for which the consequences would be the same. The quasi-singularity would happen if the sensitivity coefficients are linked for all values of the independent variable (time here). This case happens most of the time, the rank of $\boldsymbol{S}^{*t}\boldsymbol{S}^*$ is not zero but its determinant is low and its condition number built with the ratio of extreme eigenvalues (see section :

$$\text{cond}\left(\boldsymbol{S}^{*t}\boldsymbol{S}^*\right)=\frac{\lambda_{max}\left(\boldsymbol{S}^{*t}\boldsymbol{S}^*\right)}{\lambda_{min}\left(\boldsymbol{S}^{*t}\boldsymbol{S}^*\right)} \qquad \textbf{(3. 48)}$$

takes high values.

Another explanation stems from linear algebra arguments: one can consider that each sensitivity coefficient $\boldsymbol{s}_j$, that forms a (m × 1) matrix, a so-called 'colum-vector', is the components of a real vector $\vec{\boldsymbol{s}}_j$ in a m-dimensional space: the possible quasi-singularity of matrix $\boldsymbol{S}$ is caused by the fact that the vectors of the corresponding system of real vectors are 'nearly' dependent, which means that a non-zero set of n coefficients exists that makes the corresponding linear combination of these real vectors 'nearly' equal to zero (the interested reader can refer to lecture L7 of this series). Of course, the term 'nearly' needs to be quantified, that is that either all the sensitivity coefficients must have the same physical units or this analysis must be made using reduced sensitivity coefficients otherwise (see section 3.1 further down).

'Visual' and 'quantitative' criteria will now be illustrated. We introduce first the reduced sensitivity matrix $\boldsymbol{s}^*$, that enables to compare the sensitivity coefficients between themselves and to compute a covariance matrix associated to *relative* estimations (and then to compute directly relative standard deviation associated to each parameter).

3.1. The reduced sensitivity matrix $\boldsymbol{S}^*$

It is given by:

$$\boldsymbol{S}^* = \boldsymbol{S}\ \mathrm{diag}(\boldsymbol{x}) \tag{3. 49}$$

$$\text{with } \mathrm{diag}(\boldsymbol{x}) = \begin{pmatrix} x_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_n \end{pmatrix} \tag{3.50}$$

It is built with the reduced sensitivity coefficients that are defined as:

$$S_k^*(t,\boldsymbol{x}) = x_k S_k(t,\boldsymbol{x}) = x_k \left.\frac{\partial y_{mo}(t,\boldsymbol{x})}{\partial x_k}\right|_{t,x_j \text{ for } j\neq k} = \frac{\partial y_{mo}(t,\boldsymbol{x})}{\dfrac{\partial x_k}{x_k}}\Big|_{t,x_j \text{ for } j\neq k} \tag{3.51}$$

Equation (3.51) shows that the reduced sensitivity S_k^* represents the *absolute* variation of model $\partial y_{mo}(t,\boldsymbol{x})$ due to a *relative* variation of parameter $\partial x_k / x_k$. They can be also considered as the sensitivity coefficients with respect to the natural logarithm of each parameter. These reduced sensitivity coefficients have then the same unit as both model output y_{mo} and standard deviation σ_ε of the measurement noise. If their magnitude is lower than the magnitude of the measurement noise σ_ε, it means that the influence of the considered parameter on the model response will not be measurable with a correct accuracy. Consequently, the estimation of this parameter through the use of experimental measurements, if it is possible, will be highly inaccurate. Rapid information may then be given by comparing the magnitude of each reduced sensitivity coefficient to the magnitude of the measurement noise, with respect to the independent variable (here time).

In the preceding example, we have then (with n = 2 parameters):

$$\boldsymbol{S}^* = \begin{bmatrix} S_1^*(t_1) & S_2^*(t_1) \dots S_n^*(t_1) \\ \vdots & \vdots \quad \vdots \\ S_1^*(t_i) & S_2^*(t_i) \dots S_n^*(t_i) \\ \vdots & \vdots \quad \vdots \\ S_1^*(t_m) & S_2^*(t_m) \dots S_n^*(t_m) \end{bmatrix} = \begin{bmatrix} S_1^*(t_1) & S_2^*(t_1) \\ \vdots & \vdots \\ S_1^*(t_i) & S_2^*(t_i) \\ \vdots & \vdots \\ S_1^*(t_m) & S_2^*(t_m) \end{bmatrix} = \begin{bmatrix} x_1 t_1 & x_2 \\ \vdots & \vdots \\ x_1 t_i & x_2 \\ \vdots & \vdots \\ x_1 t_m & x_2 \end{bmatrix} \tag{3.52}$$

Let us notice that all the coefficients defining $\boldsymbol{x}$ have to be chosen in order to calculate (and compare) the reduced sensitivity coefficients: contrary to the sensitivity coefficients of a linear model, they do depend on the value of the parameter vector $\boldsymbol{x}$. That is why a 'nominal' value for this vector is used for this calculation, that is a value that is a priori expected to be close to its exact value in a parameter estimation problem.

3.1.1. Graphical analysis of reduced sensitivity coefficients

As said before, when nominal values of the parameters have been chosen, it could be very instructive to plot all the reduced sensitivity coefficients composing each column of $\boldsymbol{S}^*$ in the

same graph in order to 'visually' detect some future ill-conditioning of matrices $\boldsymbol{S}^{*t}\boldsymbol{S}^{*}$ (and consequently of $\boldsymbol{S}^{t}\boldsymbol{S}$) due to several factors:

- One or more columns of $\boldsymbol{S}^{*}$ have low values (in absolute value) with respect to both the other ones and to the noise level σ_ε, indicating poor sensitivities of the model to some parameters.

- Two or more column are linearly dependent, indicating correlations between some parameters that will prevent their simultaneous identification. The simplest dependence to check is the proportionality between two coefficients (see ***Figure 7*** and ***Figure 8*** for favorable and unfavorable situations). Let us note that this linear dependence has to concern the whole time interval [t_{min}, t_{max}] in order to imply an ill-conditioning of the inversion.

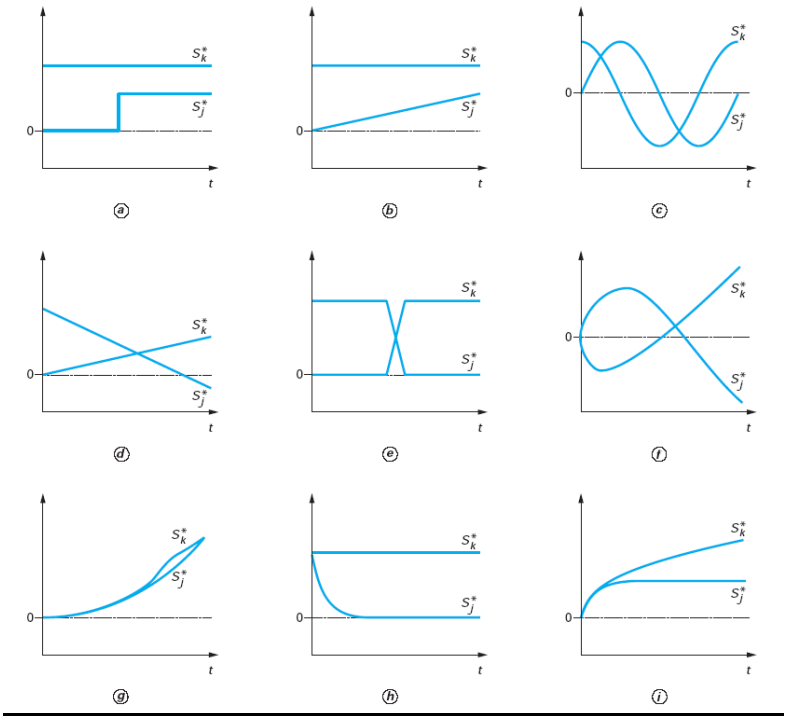


Figure 7 : some situations where reduced sensitivity coefficients S^*_k and S^*_j are linearly independent

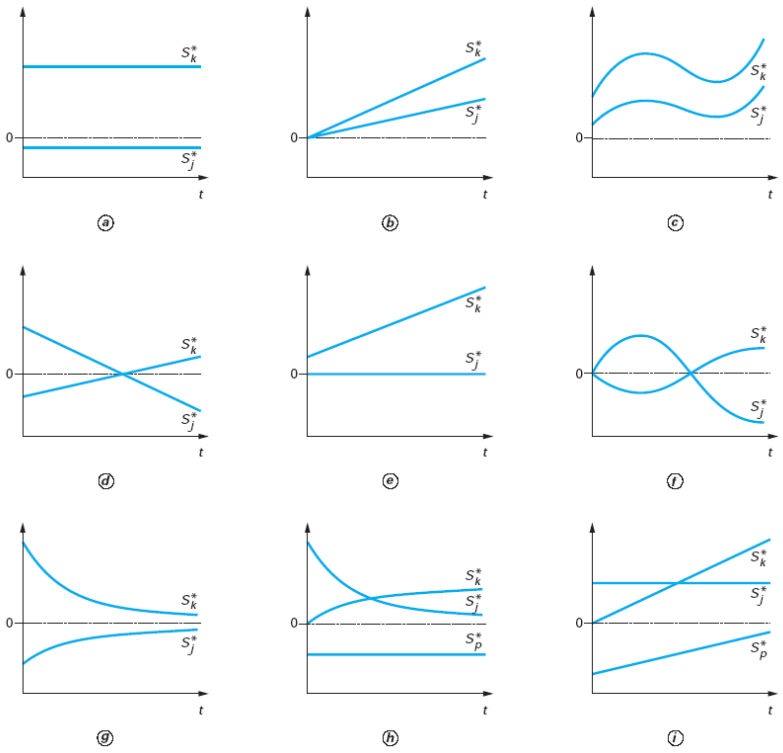


Figure 8 : some situations where reduced sensitivity coefficients S^*_k and S^*_j (and sometimes S^*_p) are linearly dependent, implying an ill-conditioning of the information matrix $S^{*t}S^*$ making it difficult, or impossible, to inverse it.

3.1.2. <u>The relative covariance matrix, and relative confidence intervals</u>

The *relative* variance-covariance matrix (size $n \times n$ for estimation of n parameters) is built the same way as the absolute variance-covariance matrix (see Eq. (3.36) and (3.39)) but the amplification matrix (inverse of the information matrix) is now built with the reduced sensitivity matrix ***S**** instead of ***S***:

$$\boldsymbol{C}^* = \sigma_\varepsilon^2 \left[\boldsymbol{S}^{*t}\boldsymbol{S}^* \right]^{-1} = \sigma_\varepsilon^2 \begin{bmatrix} \left(\frac{\sigma_1}{\hat{x}_{1,OLS}}\right)^2 & \frac{\mathrm{cov}(e_1,e_2)}{\hat{x}_{1,OLS}\hat{x}_{2,OLS}} & \cdots & \frac{\mathrm{cov}(e_1,e_n)}{\hat{x}_{1,OLS}\hat{x}_{n,OLS}} \\ & \left(\frac{\sigma_2}{\hat{x}_{2,OLS}}\right)^2 & & \frac{\mathrm{cov}(e_2,e_n)}{\hat{x}_{2,OLS}\hat{x}_{n,OLS}} \\ & & \ddots & \vdots \\ sym & & & \left(\frac{\sigma_n}{\hat{x}_{n,OLS}}\right)^2 \end{bmatrix} \qquad \textbf{(3.53)}$$

Then, the ***C**** matrix contains on its main diagonal the *n relative* variances associated to each component of the estimated vector $\hat{\boldsymbol{x}}_{OLS}$. The square root of the ith diagonal component of $\boldsymbol{C}^*$ is then the *relative* standard deviation (dimensionless) associated to the estimation $\hat{x}_{i,OLS}$ and can be expressed in %.

$$\sqrt{C_{ii}^*}(\%) = \frac{\sigma_i}{\hat{x}_{i,OLS}}, \text{ for } \quad i = 1,\ldots,n \tag{3.54}$$

Last, the half width of relative confidence interval $CI_i^{1-\alpha}(\%)$, at a level of confidence of $100(1-\alpha)\%$, associated to the estimation $\hat{x}_{i,OLS}$ (and that was evaluated with 100 stochastic simulations in Section 3.5.1.) is now given by:

$$CI_i^{1-\alpha}(\%) = t_{1-\alpha/2}(m-n) \times \sqrt{C_{ii}^*} \text{ , for } \; i=1,\ldots,n \tag{3.55}$$

Finally, the result of the estimation process of the unknown exact parameter x_i^{exact} can be presented as the following, with the *relative* confidence interval:

'x_i^{exact} has 95% chance of being in the interval $\left[\hat{x}_{i,OLS} - CI_i^{0.95}(\%) \quad \hat{x}_{i,OLS} + CI_i^{0.95}(\%)\right]$'

or : '$x_i^{exact} = \hat{x}_{i,OLS} \pm CI_i^{0.95}(\%)$ with 95% chance'

The elliptical *relative* confidence region corresponding to the scattering of estimations of ***Figure 4*** can also be computed with the *relative* information matrix $\boldsymbol{S}^{*t}\boldsymbol{S}^*$, the resulting equation expressed in the reduced coordinates $\boldsymbol{\delta x}^*$ is:

$$\begin{aligned} &\boldsymbol{\delta x}^{*t}.\boldsymbol{S}^{*t}\boldsymbol{S}^*.\boldsymbol{\delta x}^* = \Delta^2 \\ &\boldsymbol{\delta x}^* = \boldsymbol{\delta x}.diag(\hat{\boldsymbol{x}}_{OLS})^{-1} \end{aligned} \tag{3.56}$$

'Absolute' and 'relative' ellipses are plotted respectively in ***Figure 10*** and ***Figure 10*** to show that they correctly predict the extent of the 100 estimations cloud.

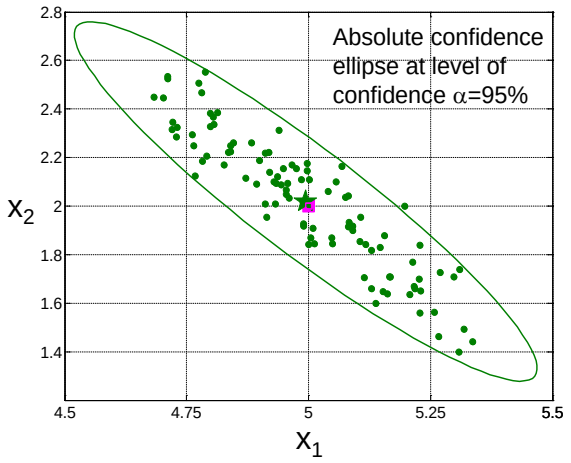


Figure 9 : 100 estimations cloud and 95% absolute confidence elliptical region around the cloud mean

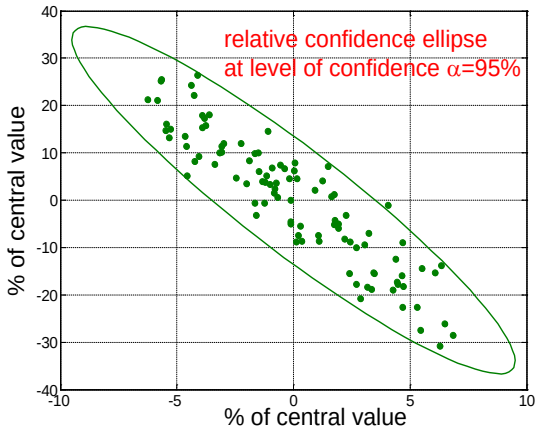


Figure 10 : 100 estimations cloud and 95% <u>relative</u> confidence elliptical region

3.2. Illustration, with a simple example, of different situations that modify the quality of estimation

In this section, the influence of some experimental parameters on the quality of estimation are illustrated for the example described in Table 2. This quality is visualized by the extent of the confidence region and some of the quantitative indicators presented above are also observed.

3.2.1. Influence of noise standard deviation σ_ε

The extension of the confidence region with respect to the standard deviation of noise measurement σ_ε, without changing its orientation, is shown in ***Figure 11***. This is conform to Eq. (3. 46) giving the ellipse area proportional to the square of σ_ε.

3.2.2. Influence of number of measurements m (in the same time range)

The extension of the 95 % confidence region with respect to the number of measurements *m*, without changing its orientation, is shown in ***Figure 12***. This is conform to Eq. (3. 46) giving the ellipse area inversely proportional to the square root of $\det(\boldsymbol{S}^t\boldsymbol{S})$, then area is inversely proportional to *m*. Then halving the noise level is better than doubling the number of measurements. This is quite obvious if one uses Eq. (3.40) and (3.43) to calculate the standard deviations and the correlation coefficient of the two OLS estimates $\hat{x}_1$ and $\hat{x}_2$:

$$\sigma_1 = \frac{\sigma}{s_t\sqrt{m}} \quad ; \quad \sigma_2 = \frac{\sigma}{\sqrt{m}}\left(1+\frac{\bar{t}^2}{s_t^2}\right)^{1/2} \quad ; \qquad \rho_{12} = -\frac{1}{(1+s_t^2/\bar{t}^2)^{1/2}} \tag{3.57}$$

where:
$$\bar{t} = \frac{1}{m}\sum_{i=1}^{m} t_i \qquad s_t^2 = \frac{1}{m}\sum_{i=1}^{m}(t_i-\bar{t})^2 \tag{3.58}$$

One clearly sees that the standard deviation of each parameter is proportional to the standard deviation of the noise and inversely proportional to the square root of the number of measurements, if the average $\bar{t}$ and the standard deviation s_t^2 of the times of measurement (Eq. (3.58)) are not changed when their number is changed.

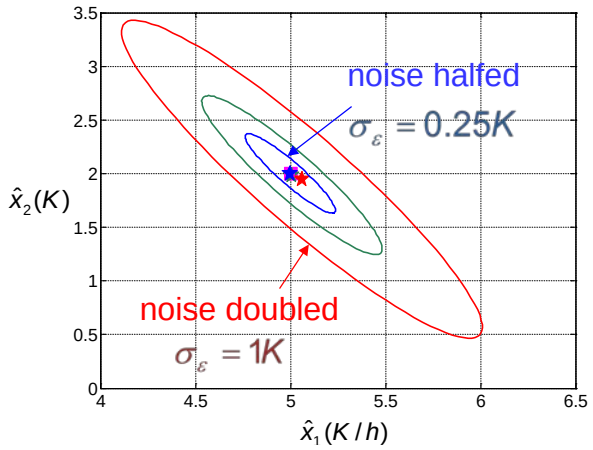


Figure 11 : Confidence ellipse extent as a function of noise level : in green (reference case) $\sigma_\varepsilon = 0.5K$ ***.***

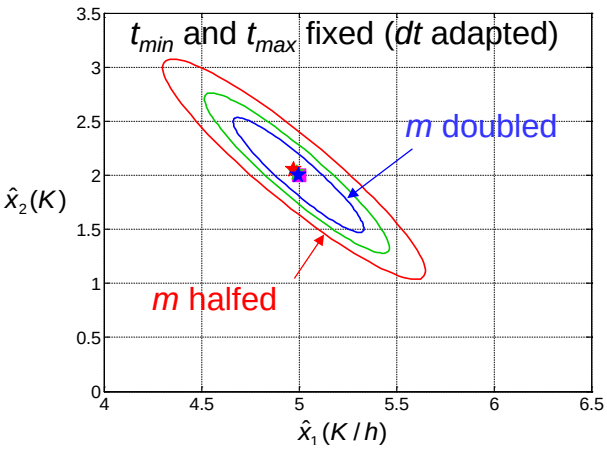


Figure 12 : Confidence ellipse extent as a function of the number of measurements m : in green (reference case) m=20.

3.2.3. Influence of time range (for m=20 measurements)

The last tested experimental factor to be varied is the time range, with a constant number of measurements (m=20), see ***Figure 13***. The results are presented in *Figure 14* and in ***Table 4***.

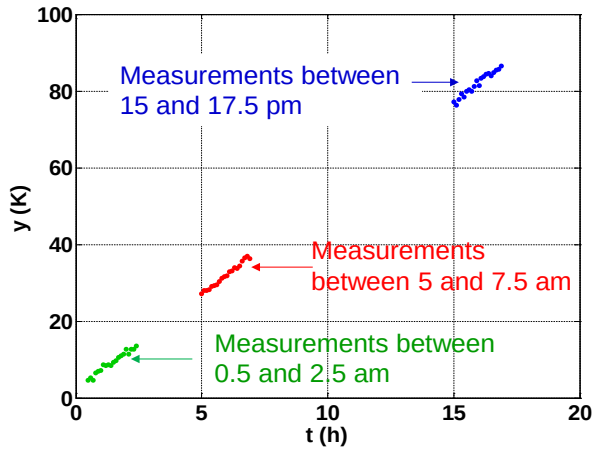


Figure 13: three time ranges are tested, giving three clouds of estimations on Figure 14.

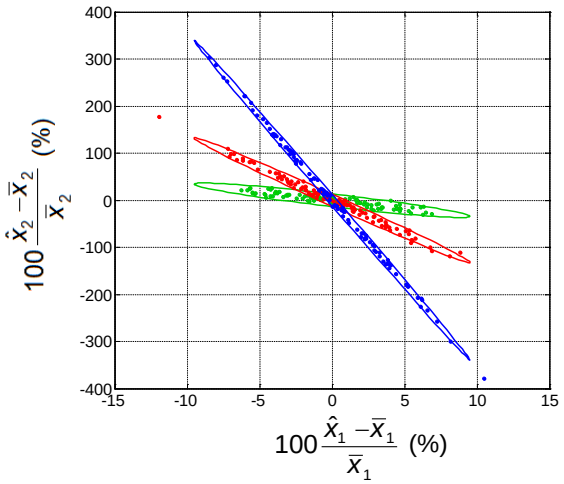


Figure 14: three clouds of estimations (corresponding to the three time ranges for the experiments) and relative 95% confidence ellipse.

Figure 14 shows that when experiments are done at 'high' time values, the confidence ellipse is growing, especially along the x_2 axis: the estimation of x_2 (intercept of the model $x_1 t + x_2$) is more and more inaccurate when the measurements are realized at high time values (far from $t = 0$). This is confirmed by the reduced sensitivity plots on ***Figure 15*** and ***Figure 16*** (see comments in legends).

Time range (h)	0.5 h -2.5 h	5 h -7.5 h	15 h -17.5 h
Central value $\bar{X}_1$ (K/h)	4.994 K/h	4.738 K/h	4.985 K/h
Absolute interval (K/h)	±0.3 K/h	±0.35 K/h	±0.35 K/h
Relative interval (%)	± 6 %	± 7 %	± 7 %
Central value $\bar{X}_2$ (K)	2.019	3.52	2.223
Absolute interval (K)	±0.5 K	±1 K/h	±5.3 K/h
Relative interval (%)	± 10 %	± 28 %	± 106 %

Table 4 : results of estimations for three different time ranges, with m=20 measurements.

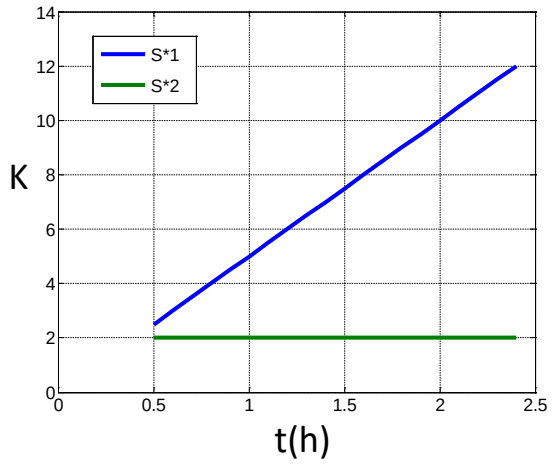


Figure 15 : first time range (between 0.5h and 2.5h). Reduced sensitivities are of same order of magnitude, sensitivity to x_1 is better than to x_2 and is increasing with time.

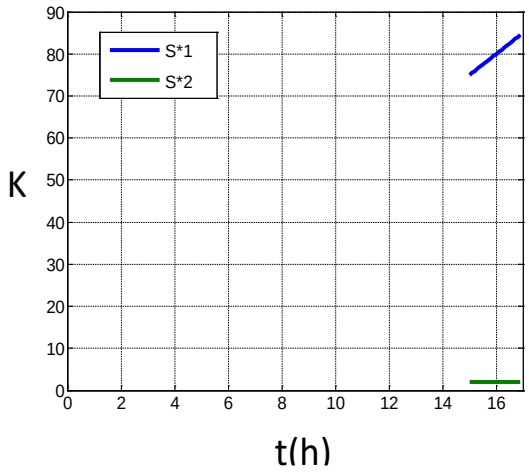


Figure 16 : third time range (between 15h and 17.5h). Reduced sensitivity to x_1 is far better than sensitivity to x_2 that appears now very close to zero comparing to S^*_1.

Last, ***Table 5*** shows multiple indicators confirming that increasing the beginning of the time range for the estimation of x_1 and x_2 is degrading the conditioning and then the quality of estimation.

Time range (h)	0.5 h -2.5 h	5 h -7.5 h	15 h -17.5 h
λ_{min} of $\boldsymbol{S}^{*t}\boldsymbol{S}^{*}$ ↑	1.03e1	6.5e-1 ↓	6.2e-2 ↓
λ_{max} of $\boldsymbol{S}^{*t}\boldsymbol{S}^{*}$ ↓	1.29e3	1.8e4 ↑	1.3e5 ↑
det($\boldsymbol{S}^{*t}\boldsymbol{S}^{*}$) ↑	1.34e4	1.18e4 ↓	8.0e3 ↓
Ellipse area ↓	3.52e-4	3.99e-4 ↑	5.9e-4 ↑
cond($\boldsymbol{S}^{*t}\boldsymbol{S}^{*}$) = λ_{max} / λ_{min} ↓	1.24e2	2.78e4 ↑↑	2.1e6 ↑↑
ρ_{12} ↓	-0.93	-0.995 ↑	-0.993 ↔

Table 5 : indicators values for the three experiments. In the first column, the arrows indicate if the indicator should be high (arrow up) or low (arrow down) to improve the conditioning.

3.3. Singular Value Decomposition of a matrix and condition number

3.3.1 Singular Value Decomposition (SVD) of a rectangular matrix

Any rectangular matrix (called $\boldsymbol{K}$ here) with real coefficients and dimension (m, n) with $m \geq n$, can be written under the form:

$$\boldsymbol{K} = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^{t}, \text{ that is } \begin{bmatrix} & \boldsymbol{K} & \end{bmatrix} = \begin{bmatrix} & \boldsymbol{U} & \end{bmatrix} \begin{bmatrix} w_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & w_n \end{bmatrix} \begin{bmatrix} & \boldsymbol{V}^{t} & \end{bmatrix} \quad \textbf{(3.59)}$$

Eq. (3.59) is sometimes called "lean" singular decomposition or "economical" SVD and involves:

- $\boldsymbol{U}$, an orthogonal matrix of dimensions (m, n) : its column vectors (the *left* singular vectors of $\boldsymbol{K}$) have a unit norm and are orthogonal by pairs : $\boldsymbol{U}^{t}\boldsymbol{U} = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the identity matrix of dimension n. Its columns are composed of the first n eigenvectors $\boldsymbol{U}_k$, ordered according to decreasing values of the eigenvalues of matrix $\boldsymbol{K}\,\boldsymbol{K}^{t}$. Let us note that, in the general case, $\boldsymbol{U}\boldsymbol{U}^{t} \neq \boldsymbol{I}_m$,

- $\boldsymbol{V}$, a square orthogonal matrix of dimensions (n, n), : $\boldsymbol{V}\,\boldsymbol{V}^{t} = \boldsymbol{V}^{t}\boldsymbol{V} = \boldsymbol{I}_n$. Its column vectors (the *right* singular vectors of $\boldsymbol{K}$), are the n eigenvectors $\boldsymbol{V}_k$, ordered according to decreasing eigenvalues, of matrix $\boldsymbol{K}^{t}\boldsymbol{K}$,

- $\boldsymbol{W}$, a square diagonal matrix of dimensions (n x n), that contains the n so-called *singular* values of matrix $\boldsymbol{K}$, ordered according to decreasing values : $w_1 \geq w_2 \geq \cdots \geq w_n$. The

singular values of matrix $\boldsymbol{K}$ are defined as the square roots of the eigenvalues of matrix $\boldsymbol{K}^t\boldsymbol{K}$. If matrix $\boldsymbol{K}$ is square and positive-definite, eigenvalues and singular values of $\boldsymbol{K}$ are the same.

Another SVD form called "Full Singular Value Decomposition" is available for matrix $\boldsymbol{K}$. In this equivalent definition, both matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ are changed: the matrix replacing $\boldsymbol{U}$ is now square (size m x m) and the matrix replacing $\boldsymbol{W}$ is now diagonal but non square (size m x n). In the case $m \geq n$, this can be written:

$$\boldsymbol{K} = \boldsymbol{U}_0\,\boldsymbol{W}_0\,\boldsymbol{V}^t \quad \text{with} \quad \boldsymbol{U}_0 = \left[\boldsymbol{U} \quad \boldsymbol{U}_{comp}\right];\ \boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{W} \\ \mathbf{0}_{(m-n)\ \mathrm{x}\ n} \end{bmatrix} \quad \text{and} \quad \dim(\boldsymbol{U}_{comp}) = m\ x\ (m - n) \tag{3.60}$$

or:

$$\begin{bmatrix} \\ \boldsymbol{K} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \boldsymbol{U} \qquad \boldsymbol{U}_{comp} \\ \\ \end{bmatrix} \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & w_n \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \\ \boldsymbol{V}^t \\ \\ \end{bmatrix} \tag{3.61}$$

Matrix $\boldsymbol{U}_{comp}$ is composed of the (m - n) left singular column vectors not present in $\boldsymbol{U}$. So, the concanated matrix $\boldsymbol{U}_0$ verifies now:

$$\boldsymbol{U}_0^t\boldsymbol{U}_0 = \boldsymbol{U}_0\boldsymbol{U}_0^t = \boldsymbol{U}\boldsymbol{U}^t + \boldsymbol{U}_{comp}\boldsymbol{U}_{comp}^t = \boldsymbol{I}_m \tag{3.62}$$

This singular value decomposition **(3.61)** can be implemented for any matrix $\boldsymbol{K}$, with real value coefficients, for $m \geq n$.

3.3.2 Interest of the Singular Value Decomposition in linear parameter estimation

We have seen above that if all the n parameters in a parameter vector $\boldsymbol{x}$ are sought for a linear model $\boldsymbol{y}_{mo}(\boldsymbol{x}) = \boldsymbol{S}\,\boldsymbol{x}$, where m noised measurements $\boldsymbol{y} = \boldsymbol{S}\,\boldsymbol{x} + \boldsymbol{\varepsilon}$ are available, and if noised $\boldsymbol{\varepsilon}$ is i.i.d., that is $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2\,\boldsymbol{I}_m$, its OLS estimator can be written:

$$\hat{\boldsymbol{x}}_{OLS} = \left(\boldsymbol{S}^t\boldsymbol{S}\right)^{-1}\boldsymbol{S}^t\,y \quad \text{with} \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{and} \quad \mathrm{cov}(\hat{\boldsymbol{x}}_{OLS}) = \sigma_\varepsilon^2\left(\boldsymbol{S}^t\boldsymbol{S}\right)^{-1} \tag{3.63}$$

The potential difficulty in its estimation may stem from the possible ill-conditioning of the square information matrix $\boldsymbol{S}^t\boldsymbol{S}$ whose inversion makes the standard deviations of its different parameters $\hat{x}_j$ become very large with respect to their exact value, see Eq. (3.53).

So, a normalized criterion can be built in order to assess the quality of the estimation of the *n* parameters.

This can be made through normalization of all the parameters x_j present in parameter vector $\boldsymbol{x}$ by a a nominal value $x_{nom,\,j}$ (which, in parameter estimation results from a prior knowledge of the order of magnitude of the corresponding parameter) to get a reduced parameter vector $\boldsymbol{x}^{red}$ without any physical unit:

$$\boldsymbol{x}^{red} = \boldsymbol{R}_{nom}^{-1}\,\boldsymbol{x} = \begin{bmatrix} x_1 / x_1^{nom} \\ x_2 / x_2^{nom} \\ \vdots \\ x_n / x_n^{nom} \end{bmatrix} \quad \text{with} \quad \boldsymbol{R}_{nom} = \text{diag}(\boldsymbol{x}^{red}) = \begin{bmatrix} x_1^{nom} & 0 & \cdots & 0 \\ 0 & x_2^{nom} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n^{nom} \end{bmatrix} \tag{3.64}$$

So the output of the linear model can be expressed in terms of the reduced sensitivity matrix $\boldsymbol{S}^*$ already presented in Section 3.1 and of the reduced (or scaled) parameter vector $\boldsymbol{x}^{red}$:

$$\boldsymbol{y}_{mo} = \boldsymbol{S}\,\boldsymbol{x} = \boldsymbol{S}\,\boldsymbol{R}_{nom}\,\boldsymbol{R}_{nom}^{-1}\,\boldsymbol{x} = \boldsymbol{S}^*\boldsymbol{x}^{red} \qquad \text{since} \quad \boldsymbol{S}^* = \boldsymbol{S}\,\boldsymbol{R}_{nom} \tag{3.65}$$

OLS estimation of this reduced parameter vector becomes, using Eq. (3.23):

$$\begin{aligned} \hat{\boldsymbol{x}}^{red} &= \boldsymbol{R}_{nom}^{-1}\,\hat{\boldsymbol{x}}_{OLS} = \boldsymbol{R}_{nom}^{-1}\,(\boldsymbol{S}^t\,\boldsymbol{S})^{-1}\boldsymbol{S}^t\,\boldsymbol{y} = \boldsymbol{R}_{nom}^{-1}\,(\boldsymbol{R}_{nom}^{-1}\,\boldsymbol{S}^{*t}\,\boldsymbol{S}^*\boldsymbol{R}_{nom}^{-1})^{-1}\,\boldsymbol{R}_{nom}^{-1}\,\boldsymbol{S}^{*t}\,\boldsymbol{y} \\ &= \boldsymbol{R}_{nom}^{-1}\,\boldsymbol{R}_{nom}(\boldsymbol{S}^{*t}\,\boldsymbol{S}^*)^{-1}\,\boldsymbol{R}_{nom}\,\boldsymbol{R}_{nom}^{-1}\,\boldsymbol{S}^{*t}\,\boldsymbol{y} = (\boldsymbol{S}^{*t}\,\boldsymbol{S}^*)^{-1}\,\boldsymbol{S}^{*t}\,\boldsymbol{y} \end{aligned} \tag{3.66}$$

And its covariance can be easily derived:

$$\text{cov}(\hat{\boldsymbol{x}}^{red}) = \sigma_\varepsilon^2\,\left(\boldsymbol{S}^{*t}\,\boldsymbol{S}^*\right)^{-1} \tag{3.67}$$

It is the same equation as Eq. (3.53). Since all the components of the reduced sensitivity matrix have the same unit as signal $\boldsymbol{y}$, and because $\boldsymbol{x}^{red}$ is dimensionless, it is possible to consider $\boldsymbol{S}^*$ as a linear application from a vector space of dimension *n* into a vector space of dimension *m*. That was not possible for the original parameter column-vector $\boldsymbol{x}$, which did not belong to a true mathematical vector space, because its coefficients had not the same units.

So, it is now possible to write the lean SVD of $\boldsymbol{S}^*$, which uses the notion of Euclidian norm of different true vectors, see Eq. (3.59):

$$\boldsymbol{S}^* = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^t \tag{3.68}$$

One can also calculate the amplification coefficient of the relative error k_r, see Eq. (1.7) in Lecture 1 of the same series:

$$k_r\,(\boldsymbol{\varepsilon}) = \frac{\left\|\boldsymbol{e}_{x\,red}\right\| \,/\, \left\|\boldsymbol{x}_{exact}^{red}\right\|}{\left\|\boldsymbol{\varepsilon}\right\| \,/\, \left\|\boldsymbol{y}_{mo}\left(\boldsymbol{x}_{exact}^{red}\right)\right\|} \quad \text{with} \quad \boldsymbol{e}_{x\,red} = \hat{\boldsymbol{x}}^{red} - \boldsymbol{x}_{exact}^{red} \tag{3.69}$$

Using the properties of matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ described above, as well as Eq. (3.66), one can show:

$$\left.\begin{array}{l} \left\| \boldsymbol{e}_{x\,red} \right\| = \left\| \boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^{t}\,\boldsymbol{\varepsilon} \right\| \le \left\| \boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^{t} \right\| \left\| \boldsymbol{\varepsilon} \right\| \\ \left\| \boldsymbol{y}_{mo}\left(\boldsymbol{x}_{exact}^{red}\right) \right\| = \left\| \boldsymbol{S}^{*}\boldsymbol{x}^{red} \right\| \le \left\| \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^{t} \right\| \left\| \boldsymbol{x}^{red} \right\| \end{array}\right\} \Rightarrow k_r\,(\boldsymbol{\varepsilon}) \le \left\| \boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^{t} \right\| \left\| \left\| \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^{t} \right\| \right\| \qquad \textbf{(3.70)}$$

One can recognize in the right-hand term of the last inequality (3.70) the product of norms of two matrices. The second matrix is simply the SVD form of the reduced sensitivity matrix $\boldsymbol{S}^{*}$ while the first one is just the pseudo inverse of $\boldsymbol{S}^{*}$, which is noted $\boldsymbol{S}^{*+}$ here.

Let us remind that the norm of any matrix $\boldsymbol{K}$ (which has not to be square) is defined by:

$$\left\| \boldsymbol{K} \right\|^{2} = \underset{\left\| \boldsymbol{z} \right\|=1}{\text{Max}} \left(\boldsymbol{z}^{t}\,\boldsymbol{K}^{t}\boldsymbol{K}\,\boldsymbol{z} \right) = w_1^2(\boldsymbol{K}) \qquad \textbf{(3.71)}$$

where $w_1(\boldsymbol{K})$ is the largest singular value of $\boldsymbol{K}$. This singular value is simply the square root of the largest (positive) value of the reduced information matrix $\lambda_{max}\left(\boldsymbol{S}^{*t}\boldsymbol{S}^{*}\right)$, see Eq. (3. 48). One can show that:

$$\left\| \boldsymbol{S}^{*} \right\| = w_1(\boldsymbol{S}^{*}) \text{ and } \quad \left\| \boldsymbol{S}^{*+} \right\| = w_1(\boldsymbol{S}^{*+}) = \frac{1}{w_n(\boldsymbol{S}^{*})} \qquad \textbf{(3.72)}$$

So, it can be shown, using Eq. (3.69), (3.70) and (3.72) that the maximum value of the amplification coefficient of the relative error k_r, that is the criterion that assesses the ill-posed character of the OLS parameter estimation problem is equal to the condition number, noted cond (.) here, of the reduced sensitivity matrix:

$$k_r\,(\boldsymbol{\varepsilon}) \le \text{cond}(\boldsymbol{S}^{*}) = \frac{w_1(\boldsymbol{S}^{*})}{w_n(\boldsymbol{S}^{*})} \qquad \textbf{(3.73)}$$

So, this condition number, defined here with the Euclidian L_2 norm, is the pertinent criterion that can be used to measure the degree of ill-posedness of a linear parameter estimation problem, whatever the value of the noise level (for an i.i.d. noise). Since it requires the construction of the reduced sensitivity matrix, it depends on the nominal values of the parameters and can change strongly, depending on this choice, even if the problem is linear.

4. Illustration on a three parameters case

4.1. All parameters are estimated or one of them is fixed

Here are the characteristics of the new model and the experimental parameters:

$x_{1nom}, [K/\sqrt{s}]$	10
$x_{2nom}, [/]$	2
$x_{3nom}, [K\sqrt{s}]$	3
Model structure $y_{mo}(t, \boldsymbol{x})$, *[K]*	$x_1 \sqrt{t} + x_2 erfc(t) + x_3 / \sqrt{t}$
Number of measurements *m*	100
Start of time range t_{min} *[s]*	0.02
Time step dt, *[s]*	0.02
Noise standard deviation σ_ε, *[K]*	0.5

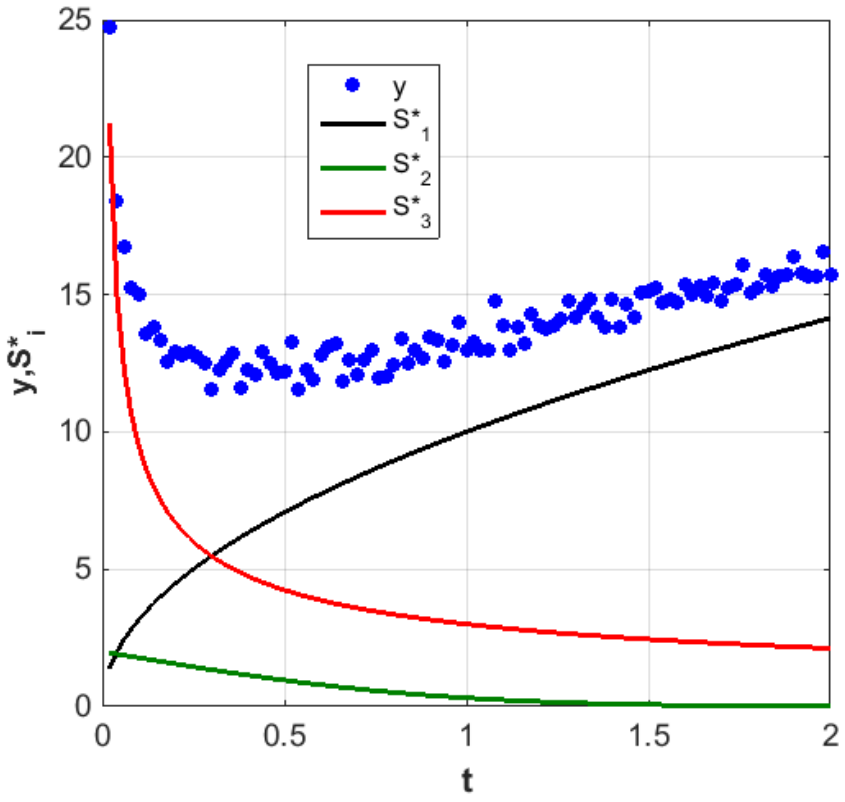


Figure 17 : **Three parameters example. Measurements (blue dot) and reduced sensitivities (at nominal values of parameters).**

Figure 18 and ***Figure 19*** shows the 100 Monte Carlo estimations of the three parameters, perfectly centred on the exact values. The condition number of $\boldsymbol{S}^{*t}\boldsymbol{S}^{*}$ here is 1325.

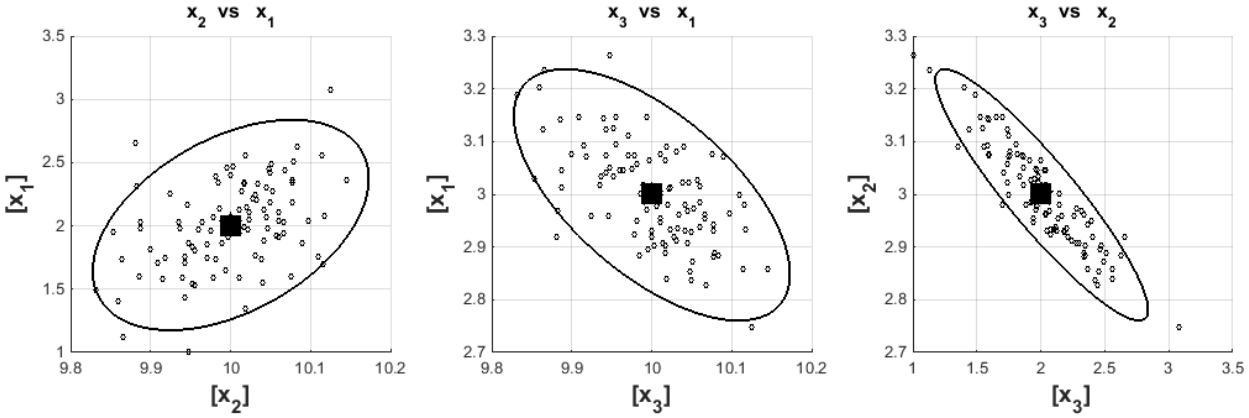


***Figure 18** : Unbiaised estimation of the three parameters: the clouds are centered on the exact value (the black square). Axes are scaled with the parameters units*

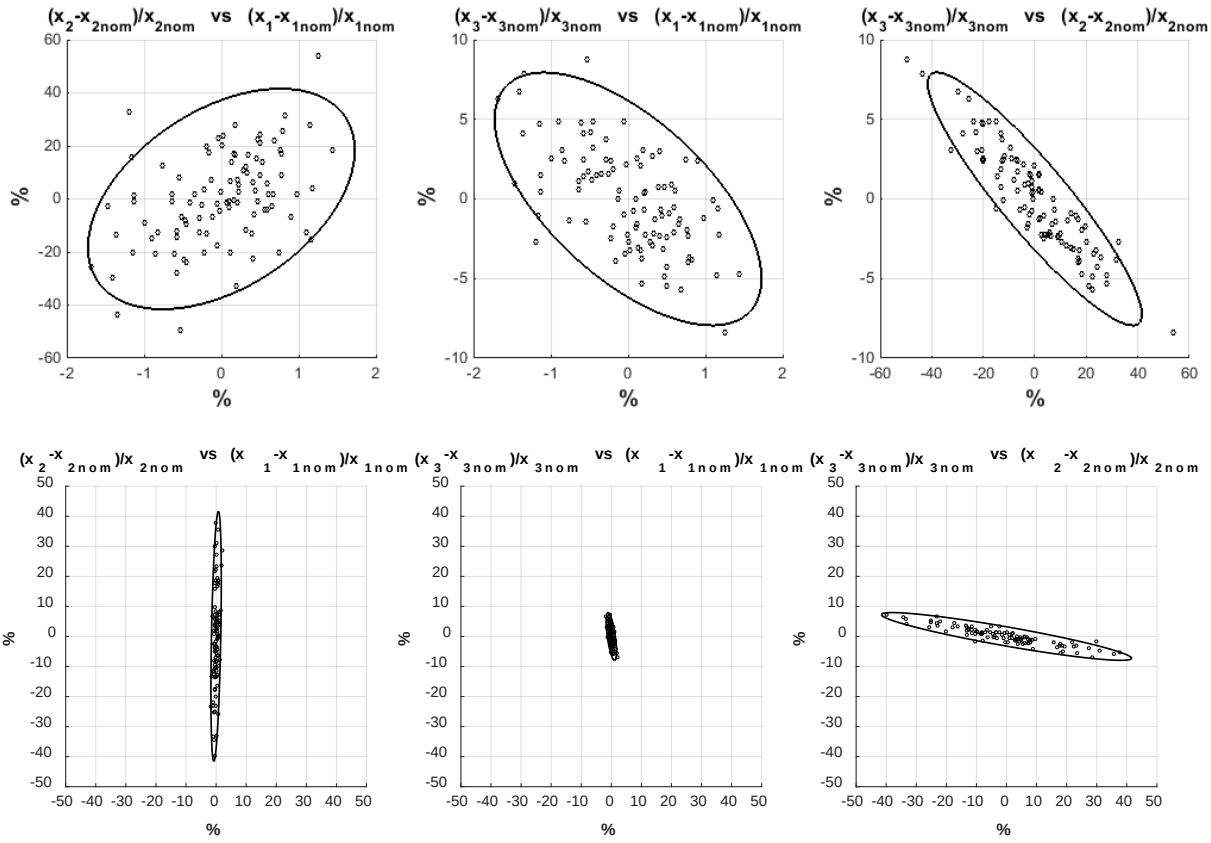


***Figure 19** : Unbiaised estimation of the three parameters. Ellipses are the relative confidence region (at level 95%) of each parameter, in axes graduated in % of the nominal values. In the second line, axes are equally graduated between -50% and 50% to visually compare the relative variance associated to each estimated parameter (dispersion of points projected on each axe) and the correlation between errors (inclination of ellipses).*

If x_1 is fixed to a wrong value (11 instead of 10, bias equal to +10%), then estimation of x_2 and x_3 is biased, see ***Figure 20*** (blue right plot). In that case, error due to fixed parameter (bias) is even higher than error due to noise measurement (variance). The condition number is better (227) for the simultaneous estimation of only 2 parameters (x_2 and x_3) but care has to be taken on the fixed value of x_1: this illustrates the 'bias-variance trade-off'.

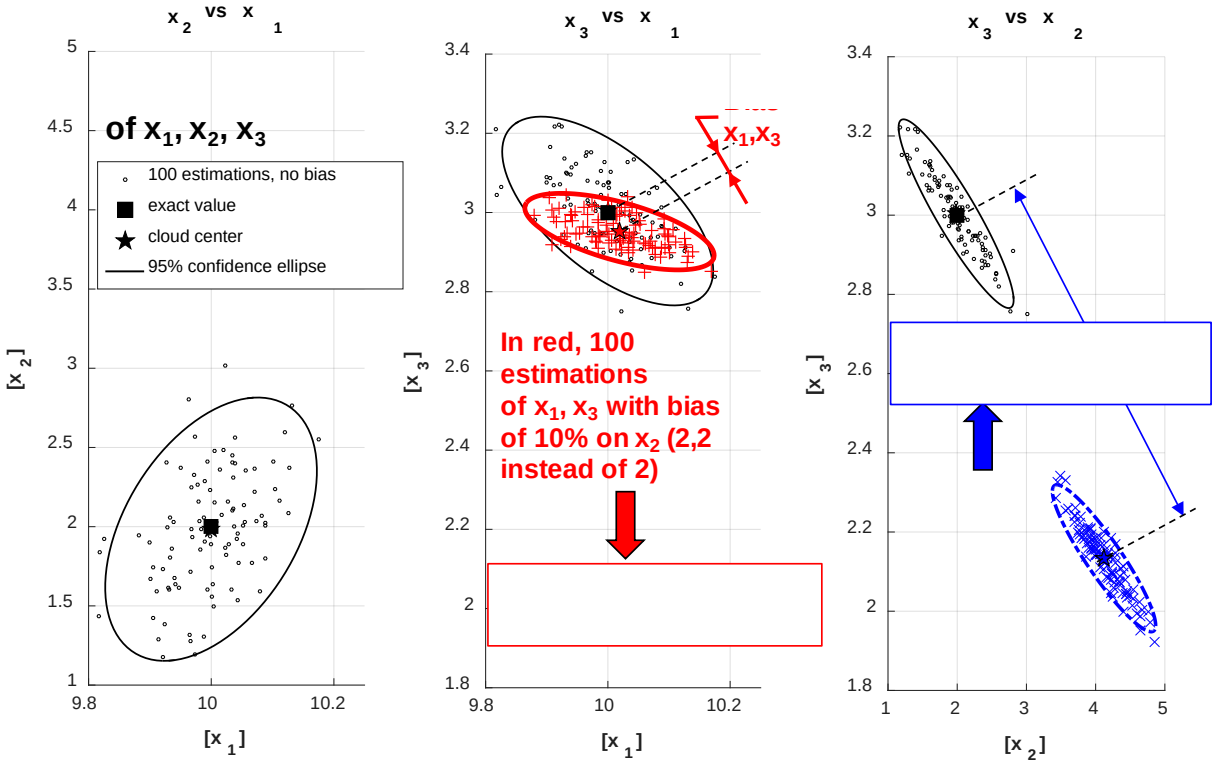


Figure 20*: *One parameter is blocked to a wrong ('biased') value, the two other are estimated (Monte Carlo run with 100 experiments). In black : no bias, same figure than Figure 17. In red (center plot), the parameter x_2 is blocked to a biased value (bias of +10%) and x_1 and x_3 are estimated and plotted in the x_3 vs x_1 plot. In blue (right plot), the parameter x_1 is blocked to a biased value (bias of +10%) and x_2 and x_3 are estimated and plotted in the x_3 vs x_2 plot. Exact values of (x_1, x_2, x_3) are (10, 2, 3). Centers of black, red and blue clouds are respectively (10.06, 1.998, 3.0003), (10.02, (2.2), 2.95) and ((11), 4.1, 2,1) where values between (.) are blocked values.

If x_2 is fixed to a wrong value (2.2 instead of 2, bias equal to +10%), then estimation of x_1 and x_3 is biased, see ***Figure 20*** (red center plot). But in that case, error due to fixed parameter (bias) is smaller than error due to noise measurement (variance). The condition number is small (equal to 9) for the simultaneous estimation of x_1 and x_3 and the amplification of bias (on x_2), given by Eq. *(3.38)* is here acceptable.

These behaviors can be related to the reduced sensitivities of ***Figure* 17**: the model is less sensitive to x_2 than x_1 during the chosen time range, then a bias on x_2 is less amplified than a bias on x_1. Last, according to Eq. (3.38), because of the inner product between $\boldsymbol{S}_r^t$ and $\boldsymbol{S}_c$ that amplifies the bias on fixed parameters $\boldsymbol{e}_c$, one has interest to block parameters whose

sensitivity coefficient $\boldsymbol{S}_c$ is the most 'orthogonal' possible to the sensitivity coefficient of estimated parameters $\boldsymbol{S}_r^t$, or in other words, the least 'collinear', or the least 'similar'.

4.2. Design of optimal experiment for a given nominal value of $\boldsymbol{x}_{nom}$

In this section, we explore the choice of experimental control variables that enable to design the experiment. In that unsteady experiment, we can change *m* the number of time of measurements and the values of these times. We add the constraints of a regular time step *dt* between each point, and the first time of acquisition being at *t=dt*. Finally, we decide also to fix the number *m* (size of the ***y*** vector of measurements), then the only variable control variable (or *design variable*) is *dt*, the associated total duration of experiment t_m being deducted by $t_m=m\ dt$. The noise measurement is also fixed to We try then to answer the question: what is the best time step *dt* (and then the best duration of experiment t_m) that will enable to estimate the parameters with the best confidence? The ***Figure 21*** shows the difference of the reduced sensitivities behavior with time for a *short* experiment (left side, $t_m=0.5s$) or a *long* experiment (right side, $t_m=4s$). Some sensitivities increase with time while others decrease, it is a potential configuration for an optimum positioning of measurement times.

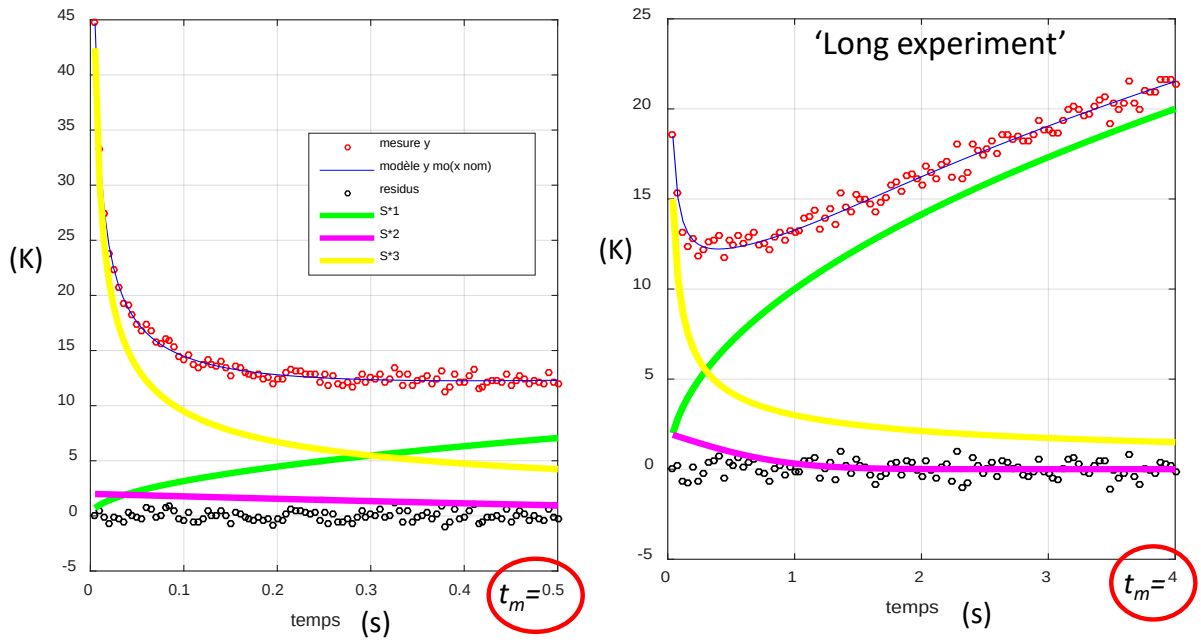


Figure 21*: Design of experiment. *With a fixed number m=100 of regularly spaced time steps beginning at $t_1=dt$*, the design variable is the time step dt (an consequently the experiment duration $t_m=m\ dt$). *Short (left) versus long (right) experiment show different measurement and sensitivities evolution with times : long experiment are more sensitive to x_1 but less sensitive to x_3 and above all less sensitive to x_2.

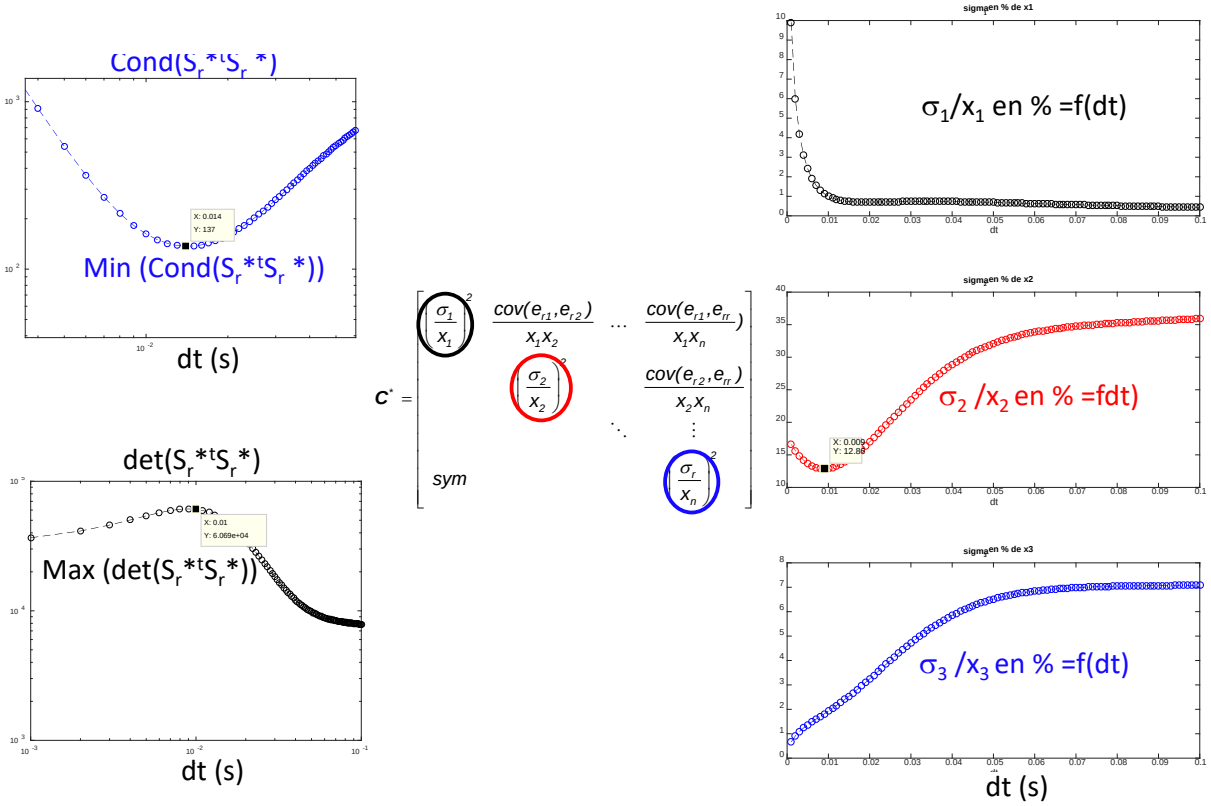


Figure 22*: Design of experiment when the 3 parameters are estimated. *Left : 2 plot of 'global' figures of merit : the conditioning of* $S^{*t}S^{*}$ to minimize *(see Eq. (3. 48)* and *(3.73))* and the determinant of $S^{*t}S^{*}$ to maximize *(see sections 2.5.4 and 3.1.2).* Right : 3 plot of 'local' *figures of merit : the three relative standard deviation (in %) associated to each estimation. They are extracted from the relative covariance matrix shown in the center (Eq. (3.53))

The ***Figure 22*** shows different way to answer the question of the best experiment if we want to estimate simultaneously the three parameters with the nominal values shown in *Figure 17*:

- With *global* indicators of the elliptical confidence region : i) the most balanced lengths of the three main axes (obtained by minimizing ***cond***($S^{*t}S^{*}$) or ii) the minimum volume of the ellipse (obtained by maximizing ***det***($S^{*t}S^{*}$)). The two criterions are optimized for a time step around *dt=0.01s* (then a duration of experiment of t_m*=1s).*

- With *local* indicators of the confidence region: the 3 relative standard deviations (that must be minimized) associated to each parameter, extracted from the relative covariance matrix given by Eq. (3.53). It confirms that confidence in estimation of x_2 and x_3 is decreasing for long experiments (and standard deviation associated to x_2 will never be better than 12% (for *dt=0.009s,* t_m*=0.9s)*. On the contrary, confidence in estimation of x_1 is increasing with long experiments, it is logical since model is more and more sensitive to x_1 at long times (see the green curve on ***Figure 21***).

If a specifical parameter is searched with a high confidence in % or if all the parameters are searched with the best confidence, the above local or global figures of merit can be considered. If the duration of experiment is a constraint and must be as short as possible, these criterions can also help to design the shortest experiment to estimate parameters with a given relative confidence in %.

5. Conclusion

The example of a linear model with respect to its two or three parameters is rich enough to introduce many tools useful in the field of parameter estimation : the sensitivity coefficients that compose the sensitivity matrix are one of these tools. This matrix has to be inverted (or the corresponding linear system of normal equations has to be solved) in the estimation problem. The variance-covariance matrix (sometimes called more simply the covariance matrix) that helps to qualify the quality of the estimation (variance of each estimation, correlation between them, size of the confidence region if the stochastic law of the measurement noise is known), uses also these coefficients. In the non linear case, the problem is often solved by assuming a local linear behaviour of the objective function to be minimized (see lecture L7 of this series).

References

Aster R. C., Borchers B., Thurber C. H., *Parameter Estimation and Inverse Problems*, Elsevier Academic Press, 2005

Beck, J. V. and Arnold K. J., *Parameter estimation in engineering and science*, John Wiley & Sons, 1977

Numerical Recipes, William Press & al., Cambridge University Press, 1989

Albert Tarentola, Inverse Problem Theory and methods for model parameter estimation, SIAM 2005

F. van der Heijden, R. P. W. Duin, D. de Ridder, D. M. J. Tax, Classification, Parameter Estimation and State Estimation, an engineering approach using Matlab®, Wiley 2004

Ne-Zheng Sun, Alexander Sun, Model Calibration and Parameter Estimation, Springer 2015

THERMAL MEASUREMENTS AND INVERSE TECHNIQUES, Edited by Helcio R. B. Orlande, Olivier Fudym, Denis Maillet, Renato M. Cotta, CRC Press, New York, ISBN : 978-1-4398-4555-4, 2011

D. Petit, D. Maillet, *Techniques inverses et estimation de paramètres (Inverse techniques and parameter estimation),* Editeur : Techniques de l'Ingénieur, Paris. Thème : Sciences Fondamentales, base : Physique-Chimie, rubrique : Mathématiques pour la physique.
- Dossier AF 4515, pp. 1- 18, janvier 2008.

- Dossier AF 4516, pp. 1-24, janvier 2008.

D. Maillet, Y. Jarny, D. Petit, *Problèmes inverses en diffusion thermique (Inverse Problems in thermal diffusion),* Editeur : Techniques de l'Ingénieur, Paris. Base documentaire: Génie Energétique.
- Dossier BE 8265 "*Modèles diffusifs, mesures et introduction à l'inversion (Diffusive models and introduction to inversion)* ", pp. 1- 54, octobre 2010,
- Dossier BE 8266 "*Formulation et résolution du problème des moindres carrés" (Formulation and solution of the least squares problem*)", pp. 1-46, janvier 2011
- Dossier BE 8267 "*Outils spécifiques de conduction inverse et de régularisation" (Specific tools for inverse conduction and regularization)*", pp. 1-46, July 2011,

Lecture 4. Measurements without contact in heat transfer: radiation thermometry

Part A: principles, implementation and pitfalls

J.-C. Krapez[1], T. Pierre[2]

[1] ONERA, The French Aerospace Lab, Salon de Provence, France
E-mail: krapez@onera.fr
[2] Univ. Bretagne Sud, UMR CNRS 6027, IRDL, F-56100, Lorient, France.
E-mail: thomas.pierre@univ-ubs.fr; philippe-le-masson@univ-ubs.fr

Abstract. The objective of this lecture is to present the main features of spectral and multispectral radiometry when applied for the purpose of temperature measurement, in particular pyrometry. The amount of thermal radiation emitted by a surface is only a fraction of the radiation emitted by a blackbody at the same temperature. The corresponding ratio is called emissivity. It is an additional unknown parameter which depends on material, wavelength, direction, and surface state. In passive radiation thermometry, whatever the number of considered wavelengths, we face an underdetermined problem, notwithstanding the fact that the atmosphere between the sensed surface and the sensor introduces itself additional unknown parameters. A series of solutions has been presented to solve the problem of emissivity and temperature separation in the field of multiwavelength pyrometry. Their performance and inherent difficulties are discussed.

List of acronyms:

- **LSMWP** Least-Squares Multi-Wavelength Pyrometry
- **MCMC** Markov Chain Monte Carlo
- **MLE** Maximum Likelihood Estimation
- **MWP** Multi-Wavelength Pyrometry
- **OLS** Ordinary Least Squares
- **RMS** Root Mean Squares
- **TES** Temperature Emissivity Separation

Scope

1. Introduction

Matter spontaneously emits electromagnetic radiation in a broad spectrum encompassing UV, visible light, infrared (IR) and microwaves. The radiance emitted by a surface depends on wavelength, temperature, direction and on the considered matter. For a solid material it also depends on the surface state, *e.g.* roughness and possibly the presence of corrosion.
Obviously, because the emitted radiance is quite sensitive to temperature, the measurement of the emitted power at a given wavelength could be used to infer the temperature. This idea is at the origin of pyrometry, thermography, and microwave radiometry.
However, the spectral radiance emitted by a material not only depends on the temperature but also on its spectral emissivity, which has thus to be known or evaluated simultaneously with the temperature. On the other hand, before reaching a remote optical sensor, the emitted radiation has been attenuated by the atmosphere. In addition, it has been combined with the radiation emitted by the atmosphere itself and the environmental radiation reflected by the aimed surface.
Evaluating the temperature from the at-sensor radiance is thus not an easy task. In this paper we present some methods that enable estimating the surface temperature. A particular emphasis is given to the temperature-emissivity separation problem.

2. Basic relations for the measured thermal radiance

2.1. Blackbody radiance

The maximum radiance emitted at given wavelength and temperature is described by Planck's law (blackbody radiance) [1]:

$$B(\lambda,T) = \frac{C_1}{\lambda^5}\left[exp\left(\frac{C_2}{\lambda T}\right) - 1\right]^{-1} \quad (1)$$

The blackbody radiance $B(\lambda,T)$ is expressed in W·m^3·sr^{-1}, wavelength λ is in m, temperature T in K, with the constants C_1 = 1.191·10^{-16} W·m^2 and C_2 = 1.439·10^{-2} m·K (notice that the blackbody radiance does not depend on direction). The blackbody radiance, as expressed by Planck's law, is described versus wavelength in Figure 1 for different temperature values (curves with a continuous line). The maximum emission is observed at a particular wavelength λ_{max} such that $exp(x_{max})\,(5 - x_{max}) = 5$, where $x_{max} \equiv C_2/\lambda_{max}T$. The solution is $x_{max} \approx 4.965$, which corresponds to $\lambda_{max}T \approx 2898\ \mu m \cdot K$ (Wien's displacement law). Hence, the peak emissive intensity shifts to shorter wavelengths as temperature rises, in inverse proportion to T.

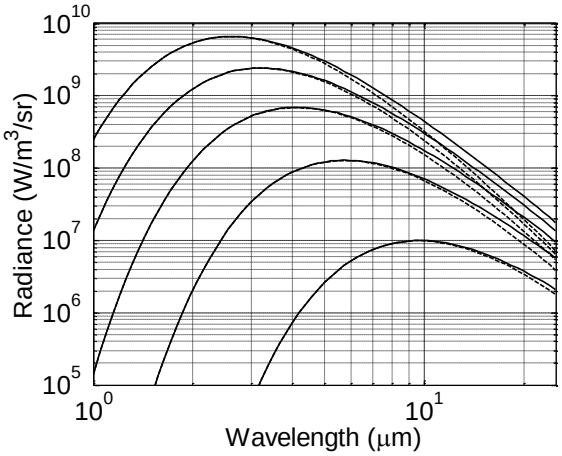


Figure 1. Blackbody radiance vs. wavelength for T = 300 K, 500 K, 700 K, 900 K and 1 100 K (from bottom to top). Planck's law: continuous lines, Wien's law: dashed lines.

A common approximation to Plank's law is Wien's law, which has been plotted in Figure 1 as well (in dashed lines):

$$B_W(\lambda,T) = \frac{C_1}{\lambda^5} exp\left(-\frac{C_2}{\lambda T}\right) \tag{2}$$

When using Wien's law, the approximation error increases with wavelength. Yet, Wien's law can be considered quite accurate in the rising part of the radiance curve. As a matter of fact, at the apex of the curve, the error has reached 0.7 % only. Also, it is less than 1 % as long as the product λT is lower than $3124\ \mu m \cdot K$.

The sensitivity of the blackbody radiance to the temperature, when considering Planck's law, is plotted in Figure 2. Figure 2-left refers to the absolute sensitivity $S = \partial B/\partial T$, whereas Figure 2-right refers to the relative sensitivity $B^{-1}\,\partial B/\partial T$. The maximum of the absolute sensitivity is observed at a wavelength λ_{Smax} such that $exp(x_{Smax})\,(6 - x_{Smax}) = 6 + x_{Smax}$ where $x_{Smax} \equiv C_2/\lambda_{Smax}T$. The solution is $x_{Smax} \approx 5.969$, which corresponds to $\lambda_{Smax}T = 2410\ \mu m \cdot K$. Notice that for a blackbody at 300 K, the maximum of radiance is observed at the wavelength $\lambda_{max} = 9.65\ \mu m$ (see Figure 1); however, the maximum sensitivity to temperature variations is observed at a shorter wavelength, namely at $\lambda_{Smax} = 8.03\ \mu m$ (see Figure 2-left). On the other hand, the *relative* sensitivity is continuously decreasing (see Figure 2-right). The asymptotic evolution is actually like $1/\lambda$ at short wavelengths. The decreasing nature of the relative sensitivity would thus favor short wavelengths for temperature measurement. However, in the meantime, the radiance progressively decreases at short wavelengths (see Figure 1). Actually, several parameters should be considered when selecting a radiative sensor together with a spectral range for temperature measurement. One should evaluate the expected radiance in the temperature range of interest, its absolute and/or relative sensitivity, together with the spectral detectivity of the candidate sensors or the corresponding noise (see *e.g.* [2]).

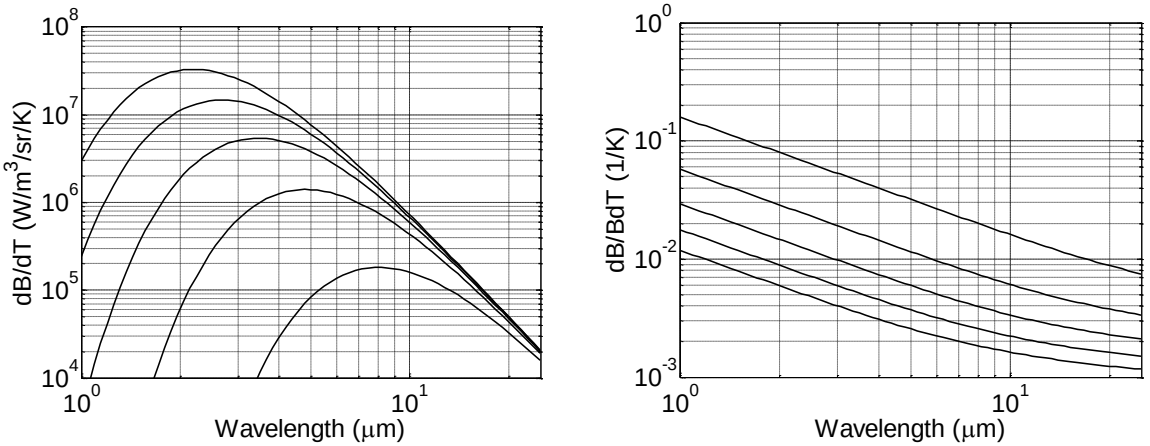


Figure 2. Absolute (left) and relative (right) sensitivity of the blackbody radiance to temperature for *T* = 300 K, 500 K, 700 K, 900 K and 1 100 K (resp., from bottom to top and top to bottom).

2.2. Emissivity and related radiative parameters

Consider a surface at temperature T and a direction defined by the zenith and azimuthal angles (θ, ϕ). The ratio between the radiance effectively emitted by the surface in this direction at wavelength λ, namely $L(\lambda, T, \theta, \phi)$, and the blackbody radiance $B_P(\lambda, T)$ at same wavelength and same temperature is called *emissivity*:

$$\varepsilon = L(\lambda, T, \theta, \phi)/B(\lambda, T)\,; \quad \varepsilon \leq 1 \tag{3}$$

Since the emissivity generally depends on wavelength and direction and since it may also depend on the surface temperature, we write it as $\varepsilon = \varepsilon(\lambda, T, \theta, \phi)$. However, if the temperature of interest is quite narrow, we may drop the T dependency for convenience and consider only $\varepsilon = \varepsilon(\lambda, \theta, \phi)$.

From the analysis of the radiation in an enclosure we can state the following relation between the emissivity and the hemispherical directional reflectance (assuming isotropic incoming radiance) [1]:

$$\varepsilon(\lambda, \theta, \phi) + \rho^{\cap'}(\lambda, \theta, \phi) = 1 \tag{4}$$

Also, the energy conservation law for an opaque material tells that the energy that is not absorbed by the surface is reflected in all directions. It leads to the following relation between the absorptivity and the directional hemispherical reflectance:

$$\alpha(\lambda, \theta, \phi) + \rho'^{\cap}(\lambda, \theta, \phi) = 1 \tag{5}$$

On the other side, the Helmholtz reciprocity principle leads to (for isotropic incoming radiance):

$$\rho^{\cap'}(\lambda, \theta, \phi) = \rho'^{\cap}(\lambda, \theta, \phi) \tag{6}$$

which, from eqs. (4) and (5), leads itself to second Kirchhoff's law, which states that the spectral emissivity in a given direction (θ, ϕ) is equal to the spectral absorptivity in the same direction:

$$\varepsilon(\lambda, \theta, \phi) = \alpha(\lambda, \theta, \phi) \tag{7}$$

2.3. Expression of the measured radiance

Assume now that an optical sensor is in the direction (θ, ϕ) to perform a measurement of the surface temperature. The radiance of the radiation leaving the surface in this direction, namely $L(\lambda, T, \theta, \phi)$, is the sum of the radiance emitted by the surface and the contribution of the radiation of radiance $L^{\downarrow}(\lambda, \theta_i, \phi_i)$ coming from the environment in all incident directions (θ_i, ϕ_i) of the upper hemisphere and then reflected by the surface (in this course, without loss of generality, we generally consider that the surface is facing up):

$$L(\lambda, T, \theta, \phi) = \varepsilon(\lambda, \theta, \phi) B(\lambda, T) + \int_{2\pi} \rho^{''}(\lambda, \theta, \phi, \theta_i, \phi_i)\, L^{\downarrow}(\lambda, \theta_i, \phi_i)\, cos\theta_i\, d\,\Omega_i \tag{8}$$

where $\rho^{''}(\lambda, \theta, \phi, \theta_i, \phi_i)$ is the bidirectional reflectance.

The radiance received by the optical sensor, which is called $\boldsymbol{L_s(\lambda, T, \theta, \phi)}$, encompasses both the radiance leaving the aimed surface and attenuated along the optical path, namely $\boldsymbol{\tau(\lambda, \theta, \phi) L(\lambda, T, \theta, \phi)}$, where $\boldsymbol{\tau(\lambda, \theta, \phi)}$ is the transmission coefficient through the air, and the radiance $\boldsymbol{L^{\uparrow}(\lambda, \theta, \phi)}$ which is self-emitted by the atmosphere along this path:

$$L_s(\lambda, T, \theta, \phi) = \tau(\lambda, \theta, \phi) L(\lambda, T, \theta, \phi) + L^{\uparrow}(\lambda, \theta, \phi) \tag{9}$$

The general radiation thermometry equation is finally:

$$L_s(\lambda, T, \theta, \phi) = \tau(\lambda, \theta, \phi) \left[\varepsilon(\lambda, \theta, \phi) B(\lambda, T) + \int_{2\pi} \rho^{''}(\lambda, \theta, \phi, \theta_i, \phi_i)\, L^{\downarrow}(\lambda, \theta_i, \phi_i)\, cos\theta_i\, d\,\Omega_i \right] + L^{\uparrow}(\lambda, \theta, \phi) \tag{10}$$

The optical sensor integrates the radiance over a narrow spectral band of width $\Delta\lambda$ centered at the wavelength λ. It delivers an electrical signal and, thanks to a calibration performed with a blackbody brought close to the sensor, a relationship can be established between this signal and the radiance $L_s(\lambda, T, \theta, \phi)$ integrated over the spectral band of width $\Delta\lambda$. Since the bandwidth $\Delta\lambda$ is small, the relationship is directly with the radiance at the wavelength λ, namely $L_s(\lambda, T, \theta, \phi)$. After proper scaling of the signal S_λ we can consider that it is a clear representation of the incoming radiance $L_s(\lambda, T, \theta, \phi)$, except it is affected by an experimental noise e_λ that for now we consider simply to be additive:

$$S_\lambda = L_s(\lambda, T, \theta, \phi) + e_\lambda \tag{11}$$

Notice that the calibration and the scaling should incorporate the contributions of the sensor optics (transmission and self-emission). Care should thus be taken that these contributions do

not change between the time interval separating the calibration process and the temperature measurements themselves.

2.4. Simplification of the radiative equation

The objective is to evaluate the surface temperature from the measurement of the radiance $L_s(\lambda, T, \theta, \phi)$ through the recording of the signal S_λ (see eq. (11)). At this point we have to deal with several unknowns: the transmission coefficient $\tau(\lambda, \theta, \phi)$ and the self-emission of the atmosphere $L^{\uparrow}(\lambda, \theta, \phi)$ along the line of sight, the hemispherical environmental radiance $L^{\downarrow}(\lambda, \theta_i, \phi_i)$, the bidirectional reflectance $\rho''(\lambda, \theta, \phi, \theta_i, \phi_i)$ for all incident directions (θ_i, ϕ_i) and the directional emissivity $\varepsilon(\lambda, \theta, \phi)$. Only when all these parameters are determined can we expect evaluating the blackbody radiance $B(\lambda, T)$ and then inferring the temperature.

A common approximation is to consider that the surface is Lambertian, *i.e.* its optical properties are direction-independent. Eq. (10) is then simplified as follows:

$$L(\lambda, T) = \varepsilon(\lambda) B(\lambda, T) + \big(1 - \varepsilon(\lambda)\big) L^{\downarrow}(\lambda) \tag{12}$$

where $L^{\downarrow}(\lambda, T)$ is the mean environmental radiance (i.e. equivalent isotropic radiance) defined according to:

$$L^{\downarrow}(\lambda, T) = \frac{1}{\pi} \int_{2\pi} L^{\downarrow}(\lambda, \theta_i, \phi_i) \, cos\theta_i \, d\,\Omega_i \tag{13}$$

We then have access to the at-sensor spectral radiance:

$$L_s(\lambda, T, \theta, \phi) = \tau(\lambda, \theta, \phi)\big[\varepsilon(\lambda) B(\lambda, T) + \big(1 - \varepsilon(\lambda)\big) L^{\downarrow}(\lambda)\big] + L^{\uparrow}(\lambda, \theta, \phi) \tag{14}$$

Generally speaking, when dealing with temperature measurement based on thermal radiation, we face two problems:

- first we have to correct the influence of the environment (reflections from nearby surfaces and from the atmosphere, along-the-path self-emission of the atmosphere and along-the-path attenuation);
- then we have to *separate emissivity and temperature*.

The atmosphere contribution through attenuation and self-emission is particularly relevant when the measurement is performed from large distances, as for example in airborne and satellite remote sensing. Specific methods for atmosphere correction have been developed for these applications. Emissivity and temperature separation methods that take advantage of the presence of the atmosphere where devised and we refer the reader to [3] for a review.
For the remaining of this presentation, we assume that an atmosphere correction has already been applied. This means, in the case of remote sensing applications, that the upwelling radiance $L^{\uparrow}(\lambda, \theta, \phi)$, the transmission coefficient $\tau(\lambda, \theta, \phi)$, and the downwelling mean radiance $L^{\downarrow}(\lambda)$ have been evaluated through simulations with a computer program designed to model atmospheric propagation of electromagnetic radiation like MODTRAN [4] or MATISSE [5].

Upon subtracting $L^{\uparrow}(\lambda,\theta,\phi)$ from the signal and then dividing by $\tau(\lambda,\theta,\phi)$ we obtain a transformed signal that is a representation of the surface-leaving radiance $L(\lambda,T,\theta,\phi)$ as expressed in eq. (8):

$$S_\lambda = \varepsilon(\lambda,\theta,\varphi)B(\lambda,T) + \int_{2\pi}\rho''(\lambda,\theta,\varphi,\theta_i,\varphi_i)L^{\downarrow}(\lambda,\theta_i,\varphi_i)cos(\theta_i)d\Omega_i + e_\lambda \quad (15)$$

where, although having been transformed, the same notations have been kept for the new signal S_λ and the corresponding noise e_λ.

In the case of Lambertian surfaces the new signal access to the surface-leaving radiance $L(\lambda,T)$ as expressed in eq. (12):

$$S_\lambda = \varepsilon(\lambda)B(\lambda,T) + [1-\varepsilon(\lambda)]L^{\downarrow}(\lambda) + e_\lambda \quad (16)$$

Notice that eq. (16) can be modified into:

$$S_\lambda = \varepsilon(\lambda)\big[B(\lambda,T) - L^{\downarrow}(\lambda)\big] + L^{\downarrow}(\lambda) + e_\lambda \quad (17)$$

2.5. Reflection component

There are different approaches for dealing with the reflection contribution, namely $\int_{2\pi}\rho^{'\,'}(\lambda,\theta,\phi,\theta_i,\phi_i)\,L^{\downarrow}(\lambda,\theta_i,\phi_i)\,cos\theta_i\,d\,\Omega_i$ in the general case or $\big(1-\varepsilon(\lambda)\big)L^{\downarrow}(\lambda)$ for lambertian surfaces.

In the case of small-scale laboratory experiments, *active pyrometry* with an additional heat source provides an efficient mean for getting rid of the reflection term. Photothermal pyrometry is an example where an additional radiative heat source is provided for the purpose of slightly heating the test material dynamically [4]-[9]. The source is either pulsed or modulated. Usually, the heat source is a laser beam aimed at the region of interest. A pyrometer is then used to measure the slight variations of the radiance (in a spectral band not including the wavelength of the radiative heat source). By considering only the variations of radiance, not the initial or DC level (as easily obtained in the modulated regime by applying lock-in detection), the contribution of the spurious reflections is eliminated since those are constant in time. Only remains a signal proportional to $\varepsilon(\lambda)\,\partial B/\partial T\,(\lambda,T)$. Furthermore, by implementing two-color pyrometry at two wavelengths λ_1 and λ_2, we can get rid of the emissivity influence (in the same way as in the static regime, as described later in § 4), and obtain a signal that depends on both $\partial B/\partial T\,(\lambda_1,T)$ and $\partial B/\partial T\,(\lambda_2,T)$ from which temperature is then easily inferred.

In remote sensing, since the downwelling mean radiance $L^{\downarrow}(\lambda)$ have already been computed with an atmospheric propagation model (in the same time as the upwelling radiance $L^{\uparrow}(\lambda,\theta,\phi)$ and the transmission coefficient $\tau(\lambda,\theta,\phi)$), the obtained value is substituted in eq. (17). The remaining unknown parameters are then the emissivity $\varepsilon(\lambda)$ and the temperature T appearing in the blackbody radiance $B(\lambda,T)$.

2.6. Introduction to the problem of temperature-emissivity separation

Whatever the configuration: active (see §2.5) or passive (see eqs. (15), (16), or (17)), radiative thermometry faces an ambiguity problem knowing that a decrease or an increase of the emissivity can be fully compensated by an increase, resp. a decrease, in temperature. Whatever the measurement wavelength, the observed signal may be explained by an infinite number of couples of emissivity values and temperature values.

It is then clear that an evaluation of the emissivity is necessary to infer the temperature from the measurement of the emitted radiance. An indirect approach consists in measuring the directional hemispherical reflectance and using eqs. (4), (5), and (6) to infer the directional emissivity. This requires using an additional radiation source and bringing close to the characterized surface an integrating hemisphere to collect all the reflected radiation. This approach was used to build several databases, which give some hints on the emissivity range and spectral variations for specific materials (see for example [10], [11], [12]).

The indirect reflectance approach is not dealt in this presentation. We rather review the approaches that consist in *simultaneously* evaluating the temperature and the emissivity, or that manage to get rid of the emissivity in the procedure of measurement of the temperature. Even though, some of the methods that are presented later can also be applied to the case described by eq. (16) or by eq. (17) in which the downwelling radiance $L^{\downarrow}(\lambda)$ is known from independent measurements or from independent simulations. We focus in the sequel on the cases where the most important contribution to the sensed signal is the surface self-emitted radiation, whereas the reflection contribution can be neglected. Pyrometry of high temperature surfaces with (relatively) cold surrounding surfaces is a typical example.

After a calibration of the optic instrument operating in a narrow spectral band around wavelength λ, we have access to the emitted radiance $L(\lambda,T)$ through the signal S_λ (albeit corrupted by a random noise e_λ):

$$S_\lambda = \varepsilon(\lambda)B(\lambda,T) + e_\lambda \tag{18}$$

In the field of pyrometry, different methods are devised depending on the number of wavelengths (*i.e.* spectral bands) used for the measurement: monochromatic pyrometry (§ 3), bispectral pyrometry (§ 4), and multiwavelength pyrometry (§ 5).

3. Single-color or monochromatic pyrometry

Let us first consider that the monochromatic measurement described in eq. (18) is errorless:

$$S_\lambda = \varepsilon(\lambda)B(\lambda,T) \tag{19}$$

An estimation of the surface emissivity then allows inferring the surface temperature. This estimation can be based on prior reflectance measurements or it can be extracted from databases. The question is then: what is the consequence of an emissivity error on the temperature evaluation?

By differentiating eq. (20) we can evaluate the sensitivity of the temperature estimation to an error on the emissivity:

$$\frac{dT}{T} = -\left(\frac{T}{B}\frac{dB}{dT}\right)^{-1}\frac{d\varepsilon}{\varepsilon} \tag{20}$$

The amplification factor $\left(\frac{T}{B}\frac{dB}{dT}\right)^{-1}$ can be easily deduced from the relative sensitivity $\frac{1}{B}\frac{dB}{dT}$ drawn in Figure 2. Also, with Wien's approximation, eq. (20) reduces to:

$$\frac{dT}{T} = -\frac{\lambda T}{C_2}\frac{d\varepsilon}{\varepsilon} \tag{21}$$

The amplification factor is about 0.08 for a temperature of 1 100 K and at 1 µm. It reaches about 0.2 for a temperature of 300 K and at 10 µm. A 10 % underestimation of emissivity thus leads to a 0.8 % overestimation of temperature in the first case (*i.e.* 8 K) and a 2 % overestimation in the second case (*i.e.* 6 K). As seen in eq. (21) the error amplification is proportional to λ. The advantage of working at short wavelength is thus evident. For this reason, some authors recommended to apply pyrometry in the visible spectrum or even in the UV spectrum (see for example [13], [14], [15]). However, although a given relative error on emissivity has a lower impact on the temperature estimation when applied at short wavelength, it should not occult the fact that a reasonable estimation of emissivity is anyway needed. The retrieved temperature is unavoidably affected by this (possibly rough) estimation of emissivity [16]. In addition, at short wavelength, both the signal and its *absolute* sensitivity to temperature decrease. The choice of the spectral range for pyrometry is thus always a compromise.

4. Two-Color pyrometry

By performing a measurement at another wavelength, we obtain new information, but unfortunately, we also introduce a new unknown, namely the spectral emissivity at this supplementary wavelength. We thus have at hand two signal values, S_1 and S_2, but three unknowns: temperature T and the two emissivity values $\varepsilon(\lambda_1)$ and $\varepsilon(\lambda_2)$. Assuming errorless signals, we have:

$$\begin{cases} S_1 = L(\lambda_1, T) = \varepsilon(\lambda_1)B(\lambda_1, T) \\ S_2 = L(\lambda_2, T) = \varepsilon(\lambda_2)B(\lambda_2, T) \end{cases} \tag{22}$$

The most popular method consists in calculating the ratio of the two spectral signals (*Ratio two-color pyrometry*):

$$R_{12} = \frac{S_1}{S_2} = \frac{\varepsilon(\lambda_1)}{\varepsilon(\lambda_2)}\frac{B(\lambda_1, T)}{B(\lambda_2, T)} = \frac{\varepsilon(\lambda_1)}{\varepsilon(\lambda_2)}\left(\frac{\lambda_2}{\lambda_1}\right)^5 \frac{exp(C_2/\lambda_2 T) - 1}{exp(C_2/\lambda_1 T) - 1} \tag{23}$$

which gives, with Wien's approximation:

$$R_{12} \approx \frac{\varepsilon(\lambda_1)}{\varepsilon(\lambda_2)}\left(\frac{\lambda_2}{\lambda_1}\right)^5 exp(-C_2/\lambda_{12}T) = \frac{\varepsilon(\lambda_1)}{\varepsilon(\lambda_2)}\left(\frac{\lambda_2}{\lambda_1}\lambda_{12}\right)^5 \frac{1}{C_1}B_W(\lambda_{12},T) \quad (24)$$

where the equivalent wavelength λ_{12} of the two-color sensor is defined by:

$$\lambda_{12}{}^{-1} = \lambda_1{}^{-1} - \lambda_2{}^{-1} \quad \Rightarrow \quad \lambda_{12} = \frac{\lambda_1\lambda_2}{\lambda_2 - \lambda_1} \quad (25)$$

Ratio two-color pyrometry thus requires knowing the emissivity ratio $\varepsilon(\lambda_1)/\varepsilon(\lambda_2)$ in order to infer temperature from the radiance ratio R_{12} according to eq. (23) or according to its approximation, eq. (24). A common assumption is that emissivity is equal at both wavelengths: $\varepsilon(\lambda_1) = \varepsilon(\lambda_2)$ (it is abusively called the *greybody* assumption even though only the two emissivity values at λ_1 and at λ_2 are required to be equal).

Like for one-color pyrometry, it is easy to relate the temperature estimation error to the emissivity error made at each wavelength:

$$\frac{dT}{T} = -\frac{\lambda_{12}T}{C_2}\left(\frac{d\varepsilon_1}{\varepsilon_1} - \frac{d\varepsilon_2}{\varepsilon_2}\right) \quad (26)$$

Let us consider these two examples defined by the triplets: [T = 1 100 K, λ_1 = 1 µm, λ_2 = 1.5 µm] and [T = 300 K, λ_1 = 10 µm, λ_2 = 12 µm]. The amplification factor reaches respectively 0.22 and 1.2. These values are 3 and 6 times higher as compared to the examples related to single-color pyrometry in the previous paragraph. The sensitivity of temperature on an error on emissivity is thus far higher with *two-color* pyrometry than with *single color* pyrometry.

The error on temperature can be lowered by reducing the equivalent wavelength λ_{12}, *i.e.* by increasing the difference between $\lambda_2{}^{-1}$ and $\lambda_1{}^{-1}$, as for example by increasing the higher wavelength λ_2 or decreasing the shorter one λ_1. In any case, the amplification factor will always be larger than the one obtained with single-color pyrometry performed at the shortest wavelength.

A common idea is that by choosing very close wavelengths, the assumption that $\varepsilon(\lambda_1) = \varepsilon(\lambda_2)$ is better justified. However, in doing so, the equivalent wavelength λ_{12} increases and the sensitivity of the radiance ratio to temperature drops dramatically. These conflicting consequences can be solved in the following way. An alternative strategy is to broaden the spectral width, more precisely to increase the $\lambda_1{}^{-1}$ - $\lambda_2{}^{-1}$ difference, (i.e. to decrease λ_{12}). Accordingly, the emissivity ratio $\varepsilon(\lambda_1)/\varepsilon(\lambda_2)$ is then likely to be far from one. A prior knowledge of the ratio $\varepsilon(\lambda_1)/\varepsilon(\lambda_2)$ is thus required for evaluating T from eq. (23) or eq. (24). If this prior estimation of the ratio $\varepsilon(\lambda_1)/\varepsilon(\lambda_2)$ is reliable, the overall benefit of this procedure is that the sensitivity of the radiance ratio to temperature is higher than before (since the equivalent wavelength λ_{12} is lower).

With single-color pyrometry performed at λ_1, the required prior knowledge is about $\varepsilon(\lambda_1)$. With (ratio) two-color pyrometry performed at λ_1 and λ_2, the required prior knowledge is about the ratio $\varepsilon(\lambda_1)/\varepsilon(\lambda_2)$. Obviously, we cannot escape the introduction of some knowledge about

emissivity. However, the advantage as compared to one-color pyrometry is that thanks to the signal ratioing, the method is insensitive to problems like a partial occultation of the line of sight, or an optical path transmission variation (provided that this transmission variation is the same in both spectral bands).

To evaluate the emissivity ratio $\varepsilon(\lambda_1)/\varepsilon(\lambda_2)$ we could resort to pyroreflectometry [17]-[19]. Each emissivity is equal to $\varepsilon(\lambda,\theta,\phi) = 1 - \rho^{\cap'}(\lambda,\theta,\phi) = 1 - \pi\eta\rho''(\lambda,\theta_i,\phi_i,\theta,\phi)$ where $\rho''(\lambda,\theta_i,\phi_i,\theta,\phi)$ is the spectral bidirectional reflectance for incident direction (θ_i,ϕ_i) and output direction (θ,ϕ), and η is a diffusion factor related to both directions. The bidirectional reflectance $\rho''(\lambda,\theta_i,\phi_i,\theta,\phi)$ is measured at both wavelengths with the use of an additional radiation source (as for example two laser beams at wavelengths λ_1 and λ_2). It is then assumed that the diffusion factor η is wavelength independent. This remaining unknown parameter is finally adjusted until the color temperatures at both wavelengths (together eventually with the ratio temperature) are made coincident. This common temperature is the true one.

In some circumstances, it may be possible to bring close to the object under study a highly reflecting surface (cold mirror). By properly choosing its shape, we obtain two benefits: first the spurious reflections from the environment are diminished, and then the apparent emissivity of the sensed surface is increased thanks to the multiple reflections of the emitted radiation between the surface and the mirror [19]. As a consequence, the temperature estimation error due to errors on emissivity now involves the ratio $\hat{\varepsilon}(\lambda_1)/\hat{\varepsilon}(\lambda_2)$ where $\hat{\varepsilon}$ is the apparent, actually amplified, emissivity (see eq. (24)). Since the ratio $\hat{\varepsilon}(\lambda_1)/\hat{\varepsilon}(\lambda_2)$ is closer to 1 the sensitivity of the temperature evaluation to the errors in emissivity estimation is therefore diminished.

Instead of evaluating the temperature from the radiance ratio in eq. (23) or eq. (24), we could get it from a least-squares minimization between the measured radiances on one side, namely S_1 at λ_1 and S_2 at λ_2 as described in eq. (18), and their theoretical counterparts on the other side. The cost function then expresses as:

$$J[T,\varepsilon(\lambda_1),\varepsilon(\lambda_2)] = [S_1 - \varepsilon(\lambda_1)B(\lambda_1,T)]^2 + [S_2 - \varepsilon(\lambda_2)B(\lambda_2,T)]^2 \tag{27}$$

and we are looking for the temperature and emissivity values that minimize this cost function, *i.e.*:

$$[T,\varepsilon(\lambda_1),\varepsilon(\lambda_2)] = \underset{T,\varepsilon(\lambda_1),\varepsilon(\lambda_2)}{argmin}\, J[T,\varepsilon(\lambda_1),\varepsilon(\lambda_2)] \tag{28}$$

This corresponds to the ordinary least squares (OLS) method, however here, the problem is underdetermined since, as said before, there are three unknown parameters: T, $\varepsilon(\lambda_1)$ and $\varepsilon(\lambda_2)$ and only two observations: S_1 and S_2. One way to solve it is to introduce a functional relationship between the two emissivity values. With this new constraint, the number of unknowns is reduced by one. An example of such a relationship is obtained by specifying a value β for their ratio:

$$\varepsilon(\lambda_1)/\varepsilon(\lambda_2) = \beta \tag{29}$$

This statement of constant emissivity-ratio is shared with the ratio method for pyrometry already invoked (see eq. (23)). We then have two methods for evaluating the temperature from

the two spectral signals S_1 and S_2: either from their ratio in eq. (23) of from the least squares equation in eqs. (27)-(28). The signals are actually corrupted by an additive random experimental noise and it is known that the expected value of the ratio is a biased estimator of the ratio of the expected values. It is thus better to use eqs. (27)-(28) for the temperature identification.

Many other functional relationships could be used. Here are a few examples:

$$\varepsilon(\lambda_1) - \varepsilon(\lambda_2) = \beta \tag{30}$$

$$1/\varepsilon(\lambda_1) - 1/\varepsilon(\lambda_2) = \beta \tag{31}$$

where β is a material-dependent constant whose value should be provided.

The emissivity compensation methods of Foley [21], Watari [22], and Anderson [23] described in [24] can all be connected to the following general relationship (32), where again β is a material-dependent constant (in [22] it was actually fixed to λ_1/λ_2).

$$\varepsilon(\lambda_1) = \varepsilon(\lambda_2)^{\beta} \tag{32}$$

The crucial point with two-color pyrometry is to find out a functional relationship like those in eq. (29) to eq. (32) together with the value of the associated parameter β. It often happens that a good choice for a given material may lead to poor results for another material or for the same material in a different state (oxidation, ageing). The great difficulty, when dealing with different materials or materials of different states, consists in finding a general functional relation capable of representing all the observed spectral variations of the emissivity.

5. Multiwavelength pyrometry

We can proceed further by adding measurements performed at additional wavelengths. In the end, we come with m values of spectral signal S_i, $i = 1,\dots,m$, which correspond to experimental measurements of m values of the spectral radiance $L(\lambda_i, T)$, $i = 1,\dots,m$. Each of these measurements is contaminated by a random error e_i, $i = 1,\dots,m$:

$$S_i = \varepsilon_i B(\lambda_i, T) + e_i \quad i = 1,\dots,m \tag{33}$$

The problem still remains underdetermined since we have at hand m equations (*i.e.* m radiance measurements), but at the same time, we face $n = m + 1$ unknowns, namely the temperature T and m values of spectral emissivity $\varepsilon_i = \varepsilon(\lambda_i)$, $i = 1,\dots,m$. The vector of parameters is called $\boldsymbol{\beta} = (\varepsilon_1 \ \dots \ \ \varepsilon_m \ T)^T$.

Multiwavelength pyrometry has been a subject of controversy for several decades [16], [25]-[52]. The experimental results showed various successes, sometimes with small temperature errors, other times with unacceptably high errors, depending on the material, on its surface state, and on the function chosen to approximate the emissivity spectrum. Even from the numerous theoretical works on this subject, it is hard to find a consensus about the advantage

or not of using many (or possibly a large number of) wavelengths [25], [26], [31], [33], [35], [36], [37], [44], [45], [48]-[53].

In the following we present few results which highlight the difficulty to obtain good and repeatable results with some multiwavelength approaches. A series of error mitigation processes are also described.

In many cases, the problem is addressed by ignoring the presence of experimental errors. As such, the system of equations to solve is:

$$S_i = \varepsilon_i B(\lambda_i, T), \quad i = 1, \ldots, m \tag{34}$$

Of course, the temperature $\hat{T}$ and emissivity values $\hat{\varepsilon}_i$, $i = 1, \ldots, m$ obtained therefrom are different from the real values T and ε_i, $i = 1, \ldots, m$ that yielded the observed signals S_i, $i = 1, \ldots, m$ (actually affected by experimental errors, see eq. (34)). This is discussed next.

We see from eq. (33) that the problem is non-linear with respect to the parameters. However, when taking the logarithm to the signal and introducing Wien's approximation to the blackbody radiance, the problem becomes linear with respect to the following transformed parameters $\boldsymbol{\beta} = \left[ln(\varepsilon_1) \ \ldots \ ln(\varepsilon_m) \ \ T_{ref}/T\right]^T$, where T_{ref} is an arbitrary reference temperature used for scaling the temperature. The new vector of observables is called $\boldsymbol{Y} = (Y_1 \ \ldots \ Y_m)^T$ and as a first approximation, we assume that the experimental error affecting the observables Y_i is additive as well (it is called e'_i):

$$Y_i \equiv ln\left(S_i \frac{{\lambda_i}^5}{C_1}\right) = ln(\varepsilon_i) - \mu_i \frac{T_{ref}}{T} + e'_i, \quad i = 1, \ldots, m \tag{35}$$

where μ_i is a constant coefficient multiplying the unknown parameter T_{ref}/T and defined by:

$$\mu_i \equiv \frac{C_2}{\lambda_i T_{ref}}, \quad i = 1, \ldots, m \tag{36}$$

5.1. Interpolation-based methods

To solve the underdetermined problem, a *potential* solution would be to *reduce by just one* the number of degrees of freedom related to the spectral emissivity data. In other words, instead of considering m unknown free parameters ε_i, $i = 1, \ldots, m$, the emissivity values ε_i should be described by a parametric function based on $m - 1$ parameters only. Several such emissivity models were proposed in the past. A polynomial of degree $m - 2$ has often been considered:

$$\varepsilon_i = \sum_{j=0}^{m-2} a_j {\lambda_i}^j, \quad i = 1, \ldots, m \tag{37}$$

The same could be done for the logarithm of emissivity when using the linearized version in eq. (35):

$$ln(\varepsilon_i) = \sum_{j=0}^{m-2} a_j \, {\lambda_i}^j, \qquad i = 1,\dots,m \tag{38}$$

In both cases, the remaining $m-1$ free parameters are the $m-1$ coefficients of the polynomial: $a_j \; j = 0,\dots,m-2$.

However, it was shown in [26], based on Wien's approximation (eq. (35)) and a polynomial representation of $ln(\varepsilon_i)$ (eq. (38)) that this method can rapidly lead to unrealistic temperature values as m increases.

Let us first assume that there is no measurement error, *i.e.* ${e_i}' = 0$, $i = 1,\dots,m$ in eq. (35):

$$Y_i = ln(\varepsilon_i) - \mu_i \frac{T_{ref}}{T}, \quad i = 1,\dots,m \tag{39}$$

Upon considering the polynomial representation of degree $m-2$ for $ln(\varepsilon_i)$ in eq. (38), the system of m equations is now based on m unknowns only. However, the introduction of the emissivity model has the consequence that the estimated temperature $\hat{T}$ obtained by solving the linear system of equations is different from the real temperature T. The estimated temperature $\hat{T}$ satisfies:

$$Y_i = \sum_{j=0}^{m-2} a_j \, {\lambda_i}^j - \mu_i \frac{T_{ref}}{\hat{T}}, \quad i = 1,\dots,m \tag{40}$$

By multiplying eq. (40) by λ_i, we obtain:

$$\lambda_i Y_i = \sum_{j=1}^{m-1} a_j \, {\lambda_i}^j - \frac{C_2}{\hat{T}}, \quad i = 1,\dots,m \tag{41}$$

which shows that $-C_2/\hat{T}$ corresponds to the constant parameter of the polynomial of degree $m-1$ interpolating the m values $\lambda_i Y_i$.

We can also notice (by subtracting eq. (41) from eq. (39) multiplied by λ_i) that the temperature error expressed through $C_2\left(1/T - 1/\hat{T}\right)$ (it has also been called "temperature correction") corresponds to the constant parameter of the polynomial of degree $m-1$ interpolating the m values $\lambda_i \, ln(\varepsilon_i)$:

$$\lambda_i \, ln(\varepsilon_i) = \sum_{j=1}^{m-1} a_j \, {\lambda_i}^j + C_2 \left(\frac{1}{T} - \frac{1}{\hat{T}}\right) \qquad i = 1,\dots,m \tag{42}$$

As a consequence, the temperature correction for single color, bicolor, and tricolor pyrometry (m = 1, 2, 3) is expressed by [16], [34]:

$$\begin{aligned} m &= 1 \qquad C_2\left(\frac{1}{T}-\frac{1}{\hat{T}}\right) = \lambda_1\, ln(\varepsilon_1) \\ m &= 2 \qquad C_2\left(\frac{1}{T}-\frac{1}{\hat{T}}\right) = \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}\, ln\left(\frac{\varepsilon_2}{\varepsilon_1}\right) \\ m &= 3 \qquad C_2\left(\frac{1}{T}-\frac{1}{\hat{T}}\right) = \frac{\lambda_1\lambda_2\lambda_3}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)}\left[\lambda_1\, ln\left(\frac{\varepsilon_2}{\varepsilon_3}\right)+\lambda_2\, ln\left(\frac{\varepsilon_3}{\varepsilon_1}\right)+\lambda_3\, ln\left(\frac{\varepsilon_1}{\varepsilon_2}\right)\right] \end{aligned} \tag{43}$$

The temperature correction involves the ratio $\varepsilon_1/\varepsilon_2$ for m = 2. With equidistant wavelengths, it involves the ratio $\varepsilon_1\varepsilon_3/{\varepsilon_2}^2$ for m = 3 and the ratio $\varepsilon_1{\varepsilon_3}^2/{\varepsilon_2}^2\varepsilon_4$ for m = 4 [34]. These ratios are of course to estimate beforehand. Assigning arbitrarily a value of 1 to the emissivity ratio for a series of metals had the consequence that the temperature estimation error increased very rapidly with the number of wavelengths [34].

It can be shown that the temperature correction limit for wavelength intervals decreasing to 0 is equal to $(-1)^{m-1}\,\lambda^m/(m-1)\,!\,d^{m-1}\,ln[\varepsilon(\lambda)]/d\,\lambda^{m-1}$ [31].

We can also recognize in eq. (42) that the temperature correction corresponds to the extrapolation at $\lambda = 0$ of the polynomial of degree $m-1$ used to interpolate the m values $\lambda_i\, ln(\varepsilon_i)$. This finding can now be developed a little more. If, *by chance*, a polynomial of degree $m-2$ could be found passing *exactly* through the m values $ln(\varepsilon_i)$, the polynomial of degree $m-1$ passing through the m values $\lambda_i\, ln(\varepsilon_i)$ would then correspond to the previous polynomial function multiplied by λ. The constant parameter (*i.e.* the temperature correction term) would thus be equal to 0. As a consequence, the estimated temperature would be the exact one. However, in reality, that a polynomial of degree $m-2$ could be found passing *exactly* through the m values $ln(\varepsilon_i)$ is highly improbable. Therefore, in practice, there is an unavoidable error regarding temperature. In addition, the error magnitude is tightly dependent on the properties of polynomial *extrapolation*. Unfortunately, it is well-known that using a polynomial interpolation to perform an extrapolation leads to increasingly high errors as the polynomial degree rises. Furthermore, things get progressively worse as the extrapolation is done far from the interpolation interval. Since the aforementioned extrapolation is done at $\lambda = 0$, this last point would actually advocate expanding the spectral range to the shortest possible wavelength (whose consequence would be to bring the extrapolation point closer to the interpolation interval), but this is only a desperate remedy.

The potentially catastrophic errors described just before are actually systematic errors, namely method errors. They are obtained even when assuming errorless spectral signals. To analyze the influence of the measurement errors, we can state, for ease, that the measurement error in channel i has the same impact as a corresponding uncertainty of the emissivity in the same channel, namely $d\varepsilon_i$. Then, the interpolation of the transformed values $\lambda_i\, ln[\varepsilon_i + d\varepsilon_i]$ leads to the same nature of extrapolation errors as described before. Finally, both extrapolation errors add together. The calculated temperature is thus more and more sensitive to measurement errors as the number of spectral bands increases.

The poor success of the *interpolation based* method originates from what has been called an *over-fitting* of the experimental data. It was finally recognized that the *interpolation based* method could be considered but only for the simpler pyrometers, actually with two or three wavelengths at most [26].

5.2. Regularization by using a low-order emissivity model

5.2.1. Emissivity models

The shortcomings of the over-fitting previously described can be mitigated by reducing the number of unknowns used to describe the emissivity spectrum.

Different models were tested in the past:

$$\varepsilon_i = \sum_{j=0}^{k} a_j \lambda_i{}^j \;\; ; \qquad i = 1,\ldots,m \;\; ; \qquad k < m-2 \quad \text{(generally } k = 1 \text{ or } 2) \tag{44}$$

$$ln(\varepsilon_i) = \sum_{j=0}^{k} a_j \lambda_i{}^j \;\; ; \qquad i = 1,\ldots,m \;\; ; \qquad k < m-2 \quad \text{(generally } k = 1 \text{ or } 2) \tag{45}$$

$$\varepsilon_i = 1/\left(1 + a_0 \lambda_i{}^2\right) \;\; ; \qquad i = 1,\ldots,m \tag{46}$$

Besides that, models of $ln(\varepsilon_i)$ based on polynomials of the variable $\lambda_i{}^{1/2}$ or $\lambda_i{}^{-1/2}$ and models involving the brightness temperature were considered in [44], [45]. A sinusoidal function of λ_i in [25], and other more "physical" models like Maxwell, Hagen-Rubens, and Edwards models were presented in [16], [38], [48].

Since the aim is merely to parameterize the m spectral values of emissivity with the help of only m_p parameters with $m_p < m-1$, there is no limit to the fertility of ideas spawned by "pyrometrists" to find new models. Indeed, new "analytical" model are constantly being published (see *e.g.* [41]-[52]), without the results being up to expectations, and for good reasons, as shown later.

On the other side, the grey-band model consists in splitting the spectrum into a small number of bands m_b, with $m_b < m$, and assigning the same emissivity value to all wavelengths λ_i belonging to a given band [33]. In this way, the number of unknowns is reduced from $m+1$ to $m_b + 1$. The bands can be narrowed to contain only three or two spectral channels as suggested in [76]. We can go even further by squeezing some bands to merely one spectral channel. The extreme limit consists in $m-1$ single-channel bands plus one dual-channel band. In that case we face a problem with m measurements and m unknowns which is thus, *in principle*, invertible. We will see that it is actually very badly conditioned.

The concept of grey-band can be generalized by allowing that the channels that are chosen to share a common emissivity value are not necessarily close together: an iterative process is described in [50] where these wavelengths are each time reshuffled according to the pseudo-continuous emissivity spectrum, *i.e.* the one defined over the m wavelengths λ_i according to:

$$\hat{\varepsilon}\left(\lambda_i, \hat{T}\right) = \frac{L(\lambda_i, T)}{B\left(\lambda_i, \hat{T}\right)} \qquad i = 1, \ldots, m \tag{47}$$

where $\hat{T}$ is the most recent temperature estimation. $\hat{\varepsilon}\left(\lambda_i, \hat{T}\right)$ is sorted from lower to higher values and the m_b bands of equal emissivity values are defined by splitting the $\hat{\varepsilon}\left(\lambda_i, \hat{T}\right)$ vector into m_b parts.

The unknown parameters of the emissivity function, together with temperature, are finally evaluated by least squares minimization. The simplest way consists in introducing Wien approximation to express the blackbody radiance and considering the observable $Y_i = ln\left[S_i \, {\lambda_i}^5 / C_1\right]$ (see eq. (35)). The logarithm of the emissivity values and the inverse of temperature (or T_{ref}/T) act as parameters of the linear model. Then, by introducing a polynomial approximation for $ln(\varepsilon_i)$ (see eq. (38)) but of degree $k < m - 2$, we come to a system of m equations:

$$Y_i = \sum_{j=0}^{k} a_j \, {\lambda_i}^j - \mu_i \frac{T_{ref}}{T} + {e'}_i, \quad i = 1, \ldots, m \tag{48}$$

and the problem now reduces to an estimation of the linear parameters a_j, $j = 0, \ldots, k$ and T_{ref}/T. This was done by a *linear least squares* method in [25], [30], [42], [43].

Otherwise, when taking for the observable the spectral signal S_i itself, we face a *non-linear least squares* problem ([27], [29], [32], [33], [35]-[41], [43]-[48], [51], [53]).

Let us add that by rearranging the m equations as described in eq. (35), we could get rid of one parameter, either a constant parameter or the temperature ([25], [30], [43]). However, it is believed that no advantage in accuracy is expected by manipulating the data to present the same information in a different form [25]. As a matter of fact, in the case of linear fitting, such a manipulation even increases the estimation error of the identified parameters.

We now consider different aspects of the Least Squares Multiwavelength Pyrometry solution (LSMWP).

5.2.2. Least-squares solution of the linearized Temperature Emissivity Separation problem (TES)

We adopt Wien's approximation and consider the vector of observables vector of observables $\boldsymbol{Y} = (Y_1 \ \ldots \ Y_m)^T$ described in eq. (35). We assume here that the experimental errors ${e'}_i$, $i = 1, \ldots, m$ are uncorrelated random variables following a Gaussian distribution of uniform variance. It is usually assumed that the spectral signal S_i, not the compound logarithm Y_i in eq. (35), is affected by a noise of uniform variance. The present approximation is valid if the spectral range is not too wide with respect to the shape of Planck's law $B(\lambda, T)$ and if the emissivity values do not span a too wide interval. Otherwise a Maximum Likelihood Estimation (MLE) is better appropriate.

According to eq. (38) where $ln(\varepsilon_i)$ is approximated by a polynomial of degree $k < m-2$, the least squares solution is:

$$\widehat{\boldsymbol{\beta}} = \left[\hat{a}_0 \quad \ldots \quad \hat{a}_k \quad \frac{1}{\hat{T}}\right]^T = \underset{a_{j,T}}{\arg\min} \sum_{i=1}^{m} \left[Y_i - \left(\sum_{j=0}^{k} a_j \lambda_i^j - \frac{C_2}{\lambda_i T} \right) \right]^2 \tag{49}$$

For numerical reasons (the reason is not only to manipulate numbers that are of similar range, but to minimize a particular condition number, see later), it is preferable to replace the wavelength λ_i in the polynomial expression by its reduced and centered value $\lambda_i *$ defined by:

$$\lambda_i *= 2\frac{\lambda_i - \lambda_{min}}{\lambda_{max} - \lambda_{min}} - 1 \tag{50}$$

In this way $\lambda_i * \in [-1,1]$. For the same reason, it is better to normalize T by T_{ref} where T_{ref} is chosen in such a way that the coefficients $\mu_i \equiv C_2/\lambda_i T_{ref}$ (see eq. (36)) are of the order of 1. The associated unknown parameter is then $\beta_T^* = T_{ref}/T$. The parameter vector is:

$$\widehat{\boldsymbol{\beta}}^* = \left[\hat{a}_0^* \quad \ldots \quad \hat{a}_k^* \quad \frac{T_{ref}}{\hat{T}}\right]^T = \underset{a_j^*,T}{\arg\min} \sum_{i=1}^{m} \left[Y_i - \left(\sum_{j=0}^{k} a_j^* \lambda_i^{*j} - \mu_i \frac{T_{ref}}{T} \right) \right]^2 \tag{51}$$

where the parameters $a_j *$ are the coefficients of the polynomial expressed in terms of $\lambda_i *$. The sensitivity matrix of this linear model is the following $m \times (k+2)$ matrix:

$$\boldsymbol{X} = \begin{bmatrix} 1 & \lambda_1 * & {\lambda_1 *}^2 & \ldots & -\mu_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & \lambda_m * & {\lambda_m *}^2 & \ldots & -\mu_m \end{bmatrix}_{m,k+2} \tag{52}$$

where the columns correspond to the sensitivity to any of the $k+2$ parameters present in vector $\boldsymbol{\beta} *$ (*i.e.* the first derivative of the model function relatively to each parameter).

The sensitivities to the first three parameters $a_j *$ ($j = 0,..2$) and to $\beta_T *$ have been plotted versus the reduced wavelength $\lambda *$ in Figure 3 for the particular case $\lambda_{max}/\lambda_{min} = 5/3$. The absolute values of the wavelength are not important, only the relative width of the total spectral band is relevant (the spectral interval [3 µm – 5 µm] satisfies the present criterion on relative width, $\lambda_{max}/\lambda_{min} = 5/3$).

The sensitivity to the first three coefficients of the model are respectively a constant, a linear function, and a quadratic function of the reduced wavelength $\lambda *$. The important question is how the sensitivity to the temperature reciprocal does compare to the former sensitivity functions? It is actually very smooth, close to linear. We thus expect a strong correlation between the parameters (since the sensitivity vectors are nearly collinear).

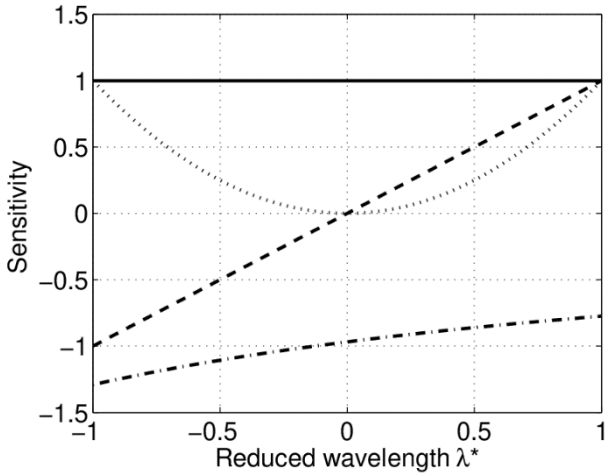


Figure 3. Sensitivity to the first three coefficients of the polynomial model (resp. continuous, dashed and dotted line) and to the inverse of the normalized temperature (dashed-dotted line). The reduced wavelength is $\lambda*$ (see eq. (50)). For this illustration, the total spectral interval is such that $\lambda_{max}/\lambda_{min}$ = 5/3.

The estimator of the parameter vector $\widehat{\boldsymbol{\beta}}*$ based on the OLS method is obtained by solving the $m \times m$ linear system (see the lecture L3 devoted to linear estimation):

$$(\boldsymbol{X}^T\boldsymbol{X})\widehat{\boldsymbol{\beta}}* = \boldsymbol{X}^T\boldsymbol{Y} \tag{53}$$

The fact that the sensitivities are nearly dependent leads to a $\boldsymbol{X}^T\boldsymbol{X}$ matrix that is near-singular. Indeed, by computing the condition number of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ (the condition number is the ratio between the maximum and minimum eigenvalues), we obtain very high values, even when the polynomial model has a low degree (see Figure 4). The condition number increases exponentially with the polynomial degree (it increases by a factor of about 100 when the polynomial degree is increased by just one). Furthermore, Figure 4 shows that increasing the number of spectral measurements in a given spectral interval brings no improvement regarding the condition number. Notice also that if the normalizations described in eqs. (49) and (50) are not applied, the condition number would reach even higher values.

The condition number describes somehow the rate at which the identified parameters changes with respect to a change in the observable; indeed, it measures the sensitivity of the solution of a system of linear equations to errors in the data. Hence, if the condition number is large, even a small error in the observables may cause a large error in the identified parameters (the condition number however only provides an upper bound). The condition number also reflects how a small change in the matrix $\boldsymbol{X}^T\boldsymbol{X}$ itself affects the identified parameters. Such a change may be due to the measurement error of the equivalent wavelength corresponding to each spectral channel. From Figure 4, a first statement is that the regularization with a polynomial model of degree 2 or higher is not efficient. But even a polynomial model of degree 1 is expected to show unstable results (the condition number is in this case of about 10^4).

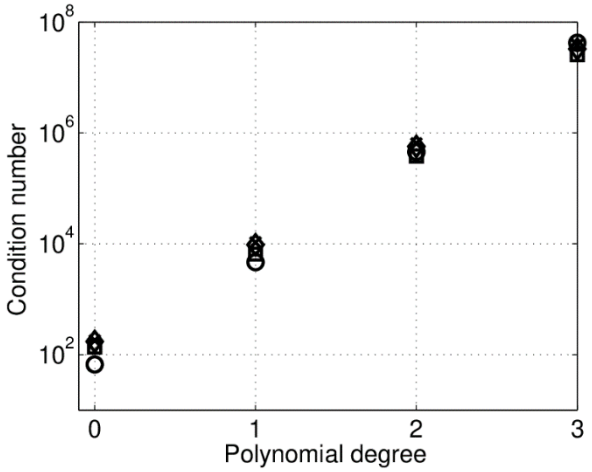


Figure 4. Condition number of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ versus the polynomial degree k of the emissivity model and for a number of spectral measurements equal to $m = k + 2$ (○), m = 7 (□), m = 30 (◊), m = 100 (x). For this illustration, the total spectral interval is such that $\lambda_{max}/\lambda_{min}$ = 5/3. The case $m = k + 2$ is the limiting case avoiding under-determination (§5.1).

The condition number has been computed in [3] for a larger spectral interval, namely for the case $\lambda_{\max}/\lambda_{\min}$ = 1.75 (the interval [8 µm – 14 µm] satisfies this criterion regarding the relative width). It was found slightly lower as compared to the present values. Increasing the relative width of the total spectral band is thus beneficial from this point of view.

However, the condition number is not all. Sometimes it could even be misleading because it only gives an upper bound of the error propagation. It is indeed better to analyze the diagonal values of the covariance matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. They actually provide the variance amplification factor for each identified parameter $P*$:

$$\left[\sigma_{\boldsymbol{\beta}*}{}^2\right] = diag((\boldsymbol{X}^T\boldsymbol{X})^{-1})\sigma^2 \tag{54}$$

where σ^2 is the variance of the observable Y_i, *i.e.* $\left(\sigma_{S_i}/S_i\right)^2$ which is here assumed independent of the spectral channel i (if instead one assumes that the radiance variance $\left(\sigma_{S_i}\right)^2$ is uniform, the result would be $\left[\sigma_{\boldsymbol{\beta}*}{}^2\right] = diag((\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1})$ where $\boldsymbol{\Psi}$ is the covariance matrix of the observable Y_i).

One should be aware that $\sigma_{\boldsymbol{\beta}*}{}^2$ merely describes the error around the mean estimator value due to the radiance error propagation to the parameters. If the mean estimator is biased, as it is the case when the true emissivity profile is not well represented by the chosen model, we should add the square *systematic error* to obtain the RMS error. The latter better represents the misfit to the true parameter value, either the temperature or a spectral emissivity value (this is described later through a Monte Carlo analysis of the inversion process).

With the polynomial model, the mean standard relative error for emissivity, which is defined by:

$$\frac{\sigma_\varepsilon}{\varepsilon} \equiv \sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{{\sigma_{\varepsilon_i}}^2}{{\varepsilon_i}^2}} \tag{55}$$

is related to the standard error of the retrieved polynomial coefficients through:

$$\frac{\sigma_\varepsilon}{\varepsilon} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left[{X_{ij}}^2\right]^T\left[{\sigma_{a_{j^*}}}^2\right]_{j=1,k}} \tag{56}$$

As such, it can be related to the uncertainty of the observable σ_Y (which corresponds to σ_S/S) through an error-amplification factor K_ε:

$$\frac{\sigma_\varepsilon}{\varepsilon} = K_\varepsilon \frac{\sigma_S}{S} \tag{57}$$

With the grey-band model, the mean standard error and the amplification factor K_ε are defined according to:

$$\frac{\sigma_\varepsilon}{\varepsilon} \equiv \sqrt{\frac{1}{m}\sum_{i=1}^{m_b}\left(\frac{\sigma_{\varepsilon_i}}{\varepsilon_i}\right)^2} = K_\varepsilon \frac{\sigma_S}{S} \tag{58}$$

From Wien's expression of the blackbody radiance, it is clear that the standard relative error for temperature is proportional to the temperature, to σ_S/S, and to a wavelength scale $\tilde{\lambda}$ representative of the spectral window (we can choose the geometric mean of the window limits: $\tilde{\lambda} \equiv \sqrt{\lambda_{min}\lambda_{max}}$). The error amplification factor for the temperature, K_T, is thus defined through:

$$\frac{\sigma_T}{T} = K_T \tilde{\lambda} T \frac{\sigma_S}{S} \tag{59}$$

The error amplification factors K_T and K_ε have been plotted in Figure 5 versus the degree of the polynomial model of emissivity, assuming again a relative bandwidth $\lambda_{max}/\lambda_{min}$ of 5/3.

A first comment is that the standard errors increase exponentially with the polynomial degree k. The rise is roughly like $exp(2k)$. The amplification factors can be reduced somewhat by widening the spectral window; in addition, the increasing rate with the polynomial degree is lower (compare with the results in [3] obtained for $\lambda_{max}/\lambda_{min}$ = 1.75).

With the grey-bands model, the standard errors increase nearly in proportion to the number of bands (see [3]).

In both cases, the standard errors decrease with the total number of spectral measurements, roughly like $m^{-1/2}$.

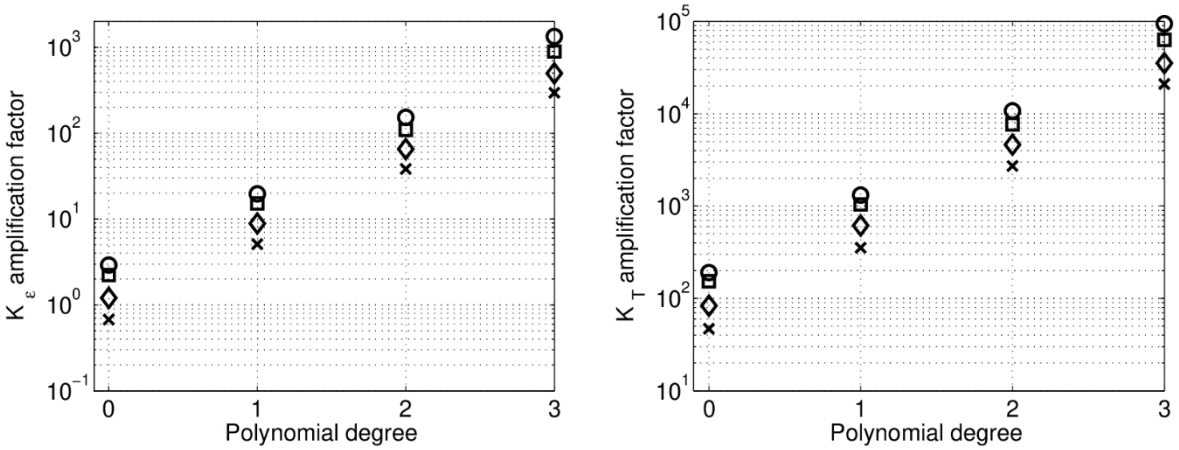


Figure 5: Left: Error amplification factor on emissivity versus the polynomial degree *m* chosen to model $ln[\varepsilon(\lambda)]$. The symbols correspond to different numbers of spectral measurements: $m = k + 2$ (○), m = 7 (□), m = 30 (◊), m =100 (x). Right: Same for the error amplification factor on temperature. The case $m = k + 2$ is the limiting case avoiding under-determination.

Regarding the bandwidth influence, let us notice that the relative error on temperature depends both on λ_{min} and λ_{max} whereas the mean relative error on emissivity only depends on the ratio $\lambda_{max}/\lambda_{min}$ (for a given value of σ_S/S in eqs. (57) and (59)).

Assuming a target at 600 K, a pyrometer with seven wavelengths between 3 µm and 5 µm and 1 % radiance noise in each spectral channel, provides temperature and emissivity values with standard errors as reported in Table 1, depending on the degree of the polynomial chosen to model the logarithm of emissivity $ln(\varepsilon_i)$.

Table 1. Polynomial model for (the logarithm of) emissivity. Root-mean square error for the estimated temperature and the emissivity depending on the degree of the polynomial model. Target temperature is 600 K. Pyrometry performed at seven wavelengths between 3 µm and 5 µm with 1 % radiance noise

Polynomial degree	σ_T (K)	σ_ε
0	2.1	0.02
1	14.5	0.13
2	107.5	1.1

The errors are already high with a linear model and they reach unacceptably high values with a polynomial model of degree two. These results seem to preclude using the least squares linear regression approach with a polynomial model of degree 2 and more.

Notice that these results have been obtained with the use of Wien's approximation. However, Planck's law is close to Wien's approximation over a large spectrum, therefore we expect that the general least squares nonlinear regression based on Planck's law also faces serious problems when using a polynomial model for emissivity.

Let us recall that the temperature and emissivity errors mentioned above only describe how the radiance errors propagate to the parameters. It has been assumed here that the emissivity spectrum otherwise *perfectly* matches the considered polynomial model. If this is not the case (which actually occurs almost every time) a *systematic error* appears and is added to the previous one. The joint errors are presented in §5.2.4 through a Monte Carlo analysis.

Applying the grey-band model to the previous example leads to the standard errors shown in Table 2. The number of grey-band can be increased up to $m_b = m - 1 = 6$ (which is the maximum to avoid underdetermination in the considered case of $m = 7$ spectral measurements).

Table 2. Grey-band model for emissivity. Root-mean square error for the estimated temperature and the emissivity depending on the number of grey-bands when assuming $m = 7$ spectral measurements. Target temperature is 600 K. Pyrometry is performed at seven wavelengths between 3 µm and 5 µm with 1 % radiance noise.

Number of bands	σ_T (K)	σ_ε
1	2.7	0.02
2	4.9	0.04
3	7.0	0.05
4	10.7	0.08
5	12.6	0.10
6	13.7	0.11

The errors increase with the number of grey-bands, starting from the values corresponding to a degree 0 polynomial and ending at values that are lower than those obtained with a polynomial of degree 1. This is interesting in the sense that even with six grey-bands, *i.e.* six degrees of freedom for emissivity, the errors do not "explode" as it is observed before by increasing the polynomial degree. The grey-bands model, although not being smooth, could thus capture more easily rapid variations in the emissivity profile like peaks.

However, as stated before, the standard errors that have been presented here only show what happens when noise corrupts the radiance emitted by a surface, but assuming that the true emissivity otherwise *perfectly follows the staircase model*. As such, with the 6-bands case, the emissivity should be *equal* in the two channels that are chosen to form the largest grey-band. As this is never strictly the case, again, a *systematic error* is added to the one shown in Table 2.

5.2.3. Another look on the solutions of the TES problem

Another way of presenting the ill-posedness of the TES problem and the difficulties in finding an appropriate regularization method consists, like in [26], in exposing first the multiple solutions to the underdetermined problem shown in eq. (34). It is clear from this set of equations that when selecting a value $\hat{T}$ for temperature, the emissivity values $\hat{\varepsilon}_i(\hat{T})$ obtained from:

$$\hat{\varepsilon}_i(\hat{T}) = S_i / B(\lambda_i, \hat{T}), \quad i = 1, \ldots, m \tag{60}$$

are such that, when combined with $\hat{T}$ they provide a *perfect* solution to the problem presented in eq. (34), namely a solution that exactly leads to the observed spectral signals. The emissivity values obtained in this way depend on the selected temperature, which explains the notation $\hat{\varepsilon}_i(\hat{T})$. Increasing the value of $\hat{T}$ entails a decrease in all spectral emissivity values and vice versa. There is an infinite number of *exact* sets of solutions $\boldsymbol{\beta} = \left[\hat{\varepsilon}_1(\hat{T}) \ \ldots \ \ \hat{\varepsilon}_m(\hat{T}) \ \ \hat{T}\right]^T$, the only limitation is that $\underset{i=1,m}{max}\, \hat{\varepsilon}_i(\hat{T}) \leq \varepsilon_{max}$ and $\varepsilon_{min} \leq \underset{i=1,m}{min}\, \hat{\varepsilon}_i(\hat{T})$. The boundary values ε_{min} and ε_{max} are chosen in accordance with the type of tested materials. Without other information ε_{max} is usually set to 1 whereas ε_{min} can be set to 0.02 since it is unusual to find surfaces with emissivities less than about 0.02, and these are very clean, polished metal surfaces [26].

As an illustration we consider a multiwavelength system operating over seven narrow spectral bands in the [3 µm – 5 µm] range, excluding the 4.3 µm CO_2 absorption band of the atmosphere. The central wavelengths are 3, 3.5, 3.7, 4, 4.6, 4.8, and 5 µm. Two hypothetic materials are considered. The first one presents an emissivity profile such that the seven emissivity values at the former seven wavelengths are distributed perfectly linearly between 0.72 at 3 µm and 0.53 at 5 µm. The emissivity of the second material has the following values: 0.72, 0.75, 0.63, 0.57, 0.56, 0.51, 0.53 at the former seven wavelengths.

These two emissivity distributions have been represented with circles in Figure 6, resp. in Figure 7. The spectral radiance is then computed at the central wavelengths according to Planck's law, assuming a temperature of 600 K. At first no measurement error is considered, it is added later. The objective is to retrieve from the seven radiance values the true temperature and the true emissivity distribution for both materials.

To start, we selected different values for the temperature $\hat{T}$ between 580 K and 660 K, and then plotted the distribution of emissivity $\hat{\varepsilon}_i(\hat{T})$ that perfectly matches with each value of $\hat{T}$ (curves with star symbols), namely that yields the same seven values of spectral radiance as those observed with the combination of true temperature and true emissivity distribution.

We notice that temperature values $\hat{T}$ as low as about 577 K could be acceptable; however, lower temperature values should be discarded since they make one of the emissivity values $\hat{\varepsilon}_i(\hat{T})$ larger than one. On the other hand, temperature values $\hat{T}$ much higher than the real temperature of 600 K could be well accepted.

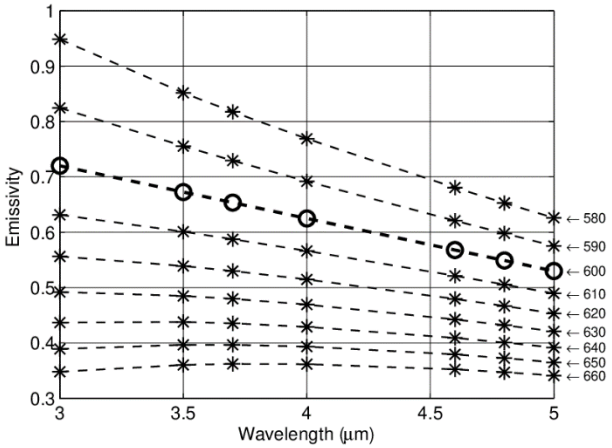


Figure 6. Emissivity profiles inferred from the spectral radiance (at seven wavelengths between 3 µm and 5 µm) by considering several hypothetical temperature values $\hat{T}$ higher or lower than the "real" temperature T = 600 K. The temperature values $\hat{T}$ are indicated on the right. The "true" emissivity distribution is with circles; it is here assumed *linear* with the wavelength.

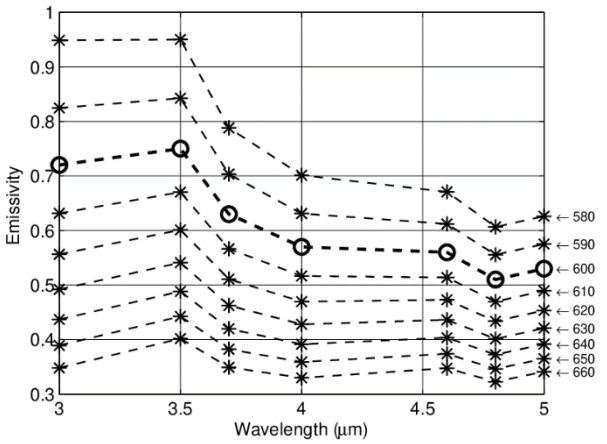


Figure 7. Same as in figure 6 for a *non-linear* emissivity distribution.

The traditional way consists in looking for a distribution of emissivity in the form of a polynomial in wavelength and performing a least squares regression on the emitted radiance. As an example, let us consider a polynomial model of degree 1. In this case, the problem can be reformulated as follows: among all hypothetic emissivity profiles represented in Figure 6 (respectively in Figure 7 for the second material), *which one is closest to a straight line*?

Let us give some indications on this notion of closeness. It is quantified by the sum of the square residues between any emissivity distribution in Figure 6 or in Figure 7 and the straight line obtained by linear regression. We are actually dealing with weighted least squares: each term should be weighted by the blackbody radiance expressed at the corresponding temperature $\hat{T}$. Hence, let us consider the weighted linear regression of a particular distribution $\hat{\varepsilon}_i(\hat{T})$; the considered weight is $B(\lambda_i, \hat{T})$. The sum of square residues is:

$$R^2(\hat{T}) = \min_{a_0,a_1} \sum_{i=1}^{m} \left[B(\lambda_i, \hat{T}) \left(\hat{\varepsilon}_i(\hat{T}) - (a_0 + a_1\lambda_i) \right) \right]^2 \tag{61}$$

Let us now consider the temperature $\hat{T}_{opt}$ for which the sum of square residues $R^2(\hat{T})$ is minimum:

$$\hat{T}_{opt} = \arg\min_{\hat{T}} \left(R^2(\hat{T}) \right) \tag{62}$$

Remember that $\hat{\varepsilon}_i(\hat{T}) B(\lambda_i, \hat{T}) = S_i$, $i = 1, \ldots, m$ for any value of $\hat{T}$ (see eq. (60)), in particular for $\hat{T}_{opt}$. Hence, we have:

$$\hat{T}_{opt} = \arg\min_{\hat{T}} \left(\min_{a_0,a_1} \sum_{i=1}^{m} \left(S_i - (a_0 + a_1\lambda_i) B(\lambda_i, \hat{T}) \right)^2 \right) \tag{63}$$

which shows that $\hat{T}_{opt}$ is also the temperature estimator obtained by the least squares minimization involving a linear emissivity model.

Notice that the previous demonstration can be extended to a polynomial model of any degree. As a consequence, when dealing with a polynomial model of degree 0, the question changes to: *which distribution $\hat{\varepsilon}_i(\hat{T})$ is closest to a horizontal line*? With a polynomial model of degree 2, it changes to: *which distribution $\hat{\varepsilon}_i(\hat{T})$ is closest to a parabola*? The demonstration can actually be extended to any other analytical model for emissivity. In the end, the general question becomes: *which distribution $\hat{\varepsilon}_i(\hat{T})$ is closest to the selected model*?

We are actually far from the aim implicitly suggested by the regression methods proposed in the literature. As a matter of fact, the emissivity models (*e.g.* polynomial functions of the wavelength) used to perform a regression of the radiance signal, give the erroneous impression that the emissivity-profile solution we are looking for is a least squares approximation of the true emissivity profile (according to the chosen model). This is absolutely not the case, as demonstrated above and illustrated next.

In Figure 6, the emissivity distribution $\hat{\varepsilon}_i(\hat{T})$ corresponding to $\hat{T}$ = 600 K is the only one to be linear. The curvature of the profiles changes depending on whether $\hat{T}$ is higher or lower than 600 K. If there is no error on the measured radiance, the best (actually perfect) match with a straight line is thus for $\hat{T}$ = 600 K, which is the right answer. Nevertheless, we have to admit that the profiles corresponding to an estimated temperature in the range 590 K < $\hat{T}$ < 610 K are very close to a straight line. It is easy to imagine that with some experimental noise added, the square residuals obtained after the linear fit would be in the same range for all profiles $\hat{\varepsilon}_i(\hat{T})$ corresponding to the former temperature range. A quantitative analysis of the noise influence is given later.

The case in Figure 7 is quite dramatic: it is evident that, among all possible solutions, the "true" profile is not the closest one to a straight line. Evidently, in this example, the distribution $\hat{\varepsilon}_i(\hat{T})$

that is closest to a straight line is obtained for a temperature $\hat{T}_{opt}$ that is much higher than the "true" value of 600 K (the profiles in the lower part in Figure 7 are indeed smoother than those in the higher part). The final solution will thus present a bias. A bias would also be obtained for the case drawn in Figure 7 if the chosen emissivity model was a polynomial of degree 0 instead of a polynomial of degree 1.

As often stated, when using LSMWP, it is necessary to choose an emissivity model that corresponds *exactly* to the true profile. The difficulty is that most often, the profile shape is unknown. A misleading thought is that LSMWP performs a fit of the true profile with the chosen model (polynomial, exponential, and so on). Actually, as seen above, performing LSMWP comes to choosing among the hypothetical solutions $\hat{\varepsilon}(\lambda,\hat{T})$, the one which fits at best to the model, in the least squares sense by weighting it with the blackbody radiance (the fit deals with ε_i if the observable is radiance and with $ln(\varepsilon_i)$ if it is the logarithm of radiance). This can lead to an emissivity profile of much higher or much lower mean value than the real one, together with an important temperature error. Actually, the problem with the classical LSMWP is that it sticks to the emissivity *shape* rather than to its *magnitude*.

5.2.4. Least squares solution of the non linear ETS problem

When using Planck's law instead of Wien's approximation, LSMWP cannot be linearized anymore. The nonlinear least squares problem can be tackled with the Levenberg-Marquardt method as provided for example by the *lsqnonlin* function from MATLAB library. When choosing a linear model for the emissivity and when the "true" emissivity profile is indeed linear this naturally leads to the right temperature and the right emissivity profile (there is no systematic error when the simulated emissivity spectrum corresponds to the chosen model). On the contrary, when the "true" emissivity profile is not linear, the identification presents a bias. For a "true" emissivity profile corresponding to the curve with circles in Figure 7, the result is reported in Figures 8 and 9. The circles in Figure 8 correspond to the theoretical radiance (no noise is added at this stage) and the stars correspond to the spectral radiance calculated from $\hat{L}(\lambda_i,\hat{T}_{opt}) = \hat{\varepsilon}_{d1}(\lambda_i)B(\lambda_i,\hat{T}_{opt})$ where $\hat{\varepsilon}_{d1}(\lambda)$ is the polynomial of degree 1 to which the distributions $\hat{\varepsilon}_i(\hat{T})$ come closest (the one which is closest is $\hat{\varepsilon}_i(\hat{T} = 652\ K)$). A perfect match for the radiance is of course impossible: the low order model chosen for emissivity (polynomial of degree 1) cannot explain the observed variations of the radiance.

The least squares procedure reveals that the $\hat{\varepsilon}_i(\hat{T})$ distribution in Figure 7 that fits at best to a straight line (considering the weighting with the blackbody radiance) is the one corresponding to the temperature $\hat{T}_{opt}$= 652 K. The stars in Figure 9 correspond to $\hat{\varepsilon}_i(\hat{T} = 652\ K)$ and the continuous curve is the unique line to which the distributions $\hat{\varepsilon}_i(\hat{T})$ come at closest, namely $\hat{\varepsilon}_{d1}(\lambda)$. The systematic error is thus +52 K for temperature and between -0.16 and -0.35 for emissivity.

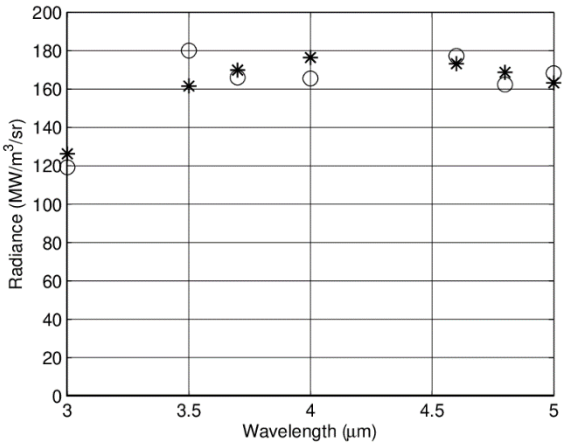


Figure 8. Inversion result for the emissivity profile represented with circles in Figure 7 when using a linear model for emissivity. Here, the circles represent the "true" noiseless radiance (true temperature: T = 600 K), the stars correspond to the emitted radiance according to the solution (*i.e.* the emissivity distribution that is closest to a straight line, which is obtained for $\hat{T}_{opt}$ = 652 K).

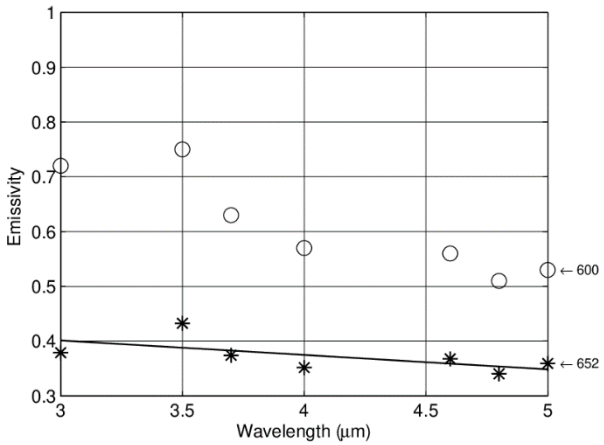


Figure 9. Inversion result for the emissivity distribution from Figure 7 when using a linear model. The "true" emissivity distribution is shown with circles (T = 600 K). The solution is represented with stars (the associated temperature $\hat{T}_{opt}$ is 652 K). The linear regression profile of the solution is represented with a continuous line ($\hat{\varepsilon}_{d1}(\lambda)$).

If the fitting happened be too far from the $\hat{\varepsilon}_i(\hat{T}_{opt})$ profile, the model should be changed. For this particular example, however, changing to a quadratic model leads to a complete failure: the profile in Figure 7 that is closest to a polynomial of degree 2 is the one corresponding to 500 K and the retrieved (hypothetical) emissivity spectrum ranges between 1.4 and 3.7! Obviously, the constraint $\hat{\varepsilon}_i(\hat{T}) < 1$ should imposed. The acceptable solution would then be the profile associated to $\hat{T}$ = 577 K which nevertheless means a 23 K underestimation.

Let us now analyze the influence of the measurement noise on the temperature and emissivity separation performance. This can be easily performed by simulating experiments where the

theoretical radiance is corrupted with artificial noise. The radiance is altered by adding values that are randomly generated with a predetermined probability density function. We assume a Gaussian distribution with a spectrally uniform standard deviation. We fix it to a value ranging from 0.2 % to 6.0 % of the maximum radiance (additive noise). The least squares minimization is performed without constraint (*i.e.* without imposing ε_i < 1) in order to highlight the mathematical (poor) stability of the inversion procedure. A series of 200 radiance spectra is treated for each noise level and for the two nominal emissivity profiles described in Figures 6 and 7. We chose again a linear emissivity model for the LSMWP inversion. The results for the maximum root mean square emissivity error among the seven channels are plotted in Figure 10-left. Those for the root mean square error on temperature are plotted in Figure 10-right. We can notice that:

- for the "true" profile of linear type (crosses), the RMS error on temperature and on emissivity increases proportionally to the radiance noise level. In particular, the RMS errors are 0.1 for emissivity and 12 K for temperature when the noise is 1 %.
- for the "true" profile of non-linear type (circles), the RMS errors are first dominated by the systematic error, which corresponds to the model implementation error (the chosen model –polynomial of degree 1 – is too crude to match the "true" profile); statistic errors due to the measurement noise dominate only when the noise is higher than 2-3 %. The RMS errors are 0.36 for emissivity and 54 K for temperature when the noise is 1 %.

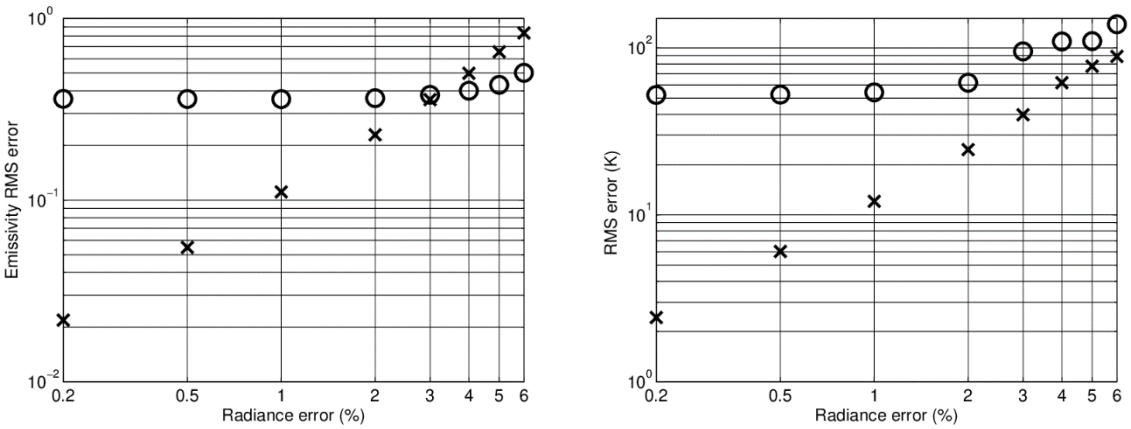


Figure 10. Statistical analysis (Monte Carlo sampling with 200 simulated experiments) of the measurement noise influence on the identified emissivity when using a linear emissivity model. The "true" emissivity was considered linear (crosses – refer to Figure 6) or non-linear (circles – refer to Figure 7). Multispectral measurement in seven channels between 3 and 5 µm. Left: emissivity error, Right: temperature error.

Let us also add that the inversion leads to a systematic error as soon as the "true" profile departs from a straight line. The previous analysis allows us to evaluate the magnitude of this error when the deviation is small. Statistically, by considering several "true" profiles close to the nominal straight line in Figure 6, the RMS of the systematic errors would be equal to the RMS of the statistic errors obtained by adding the same amount of measurement noise. For this reason, a "true" profile departing by as little as 1 % from a straight line leads to an emissivity bias whose RMS value is about 0.1. The temperature quadratic mean error is in this case about 12 K which is far from negligible. This result highlights the considerable importance of choosing

the right emissivity model. This impact can be reduced by increasing the number of spectral channels (the trend is like $N^{-1/2}$), at the condition that the departure from the profile model is randomly distributed.

As a conclusion we can state that:

- Even by reducing the number of unknowns, as it is done here by modeling the spectral emissivity with a polynomial of low degree, the problem remains badly conditioned; with a polynomial model (either for $\varepsilon(\lambda)$ or for $ln\,\varepsilon\,(\lambda)$), reasonable inversion results are expected only when the degree is 1 or 0.
- Important systematic errors appear as soon as the real emissivity departs from the considered model: 1 % departure from a straight line already leads to 12 K RMS error. More complicated spectral shapes lead to unpredictably high systematic errors (54 K for the considered example).
- Even if the real emissivity values at the sampled wavelengths *perfectly fitted* to a straight line, the demand on radiance measurement precision is very high: as a matter of fact, no more than 0.2 % noise is allowed to get an RMS error lower than 2.5 K near 600 K for a 7-band pyrometer between 3 µm and 5 µm.

Finally, LSMWP is not performing well for simultaneous evaluation of temperature and emissivity. Reasonable RMS values can be obtained only when the emissivity spectrum perfectly matches with the chosen emissivity model. Otherwise, important systematic errors are encountered. The problem is that, apart from a few exceptions, it is not known beforehand whether the emissivity of a tested material conforms to such a model or another.

The results are disappointing because the inversion is based on the emitted spectral radiance only. Good results can be obtained by taking advantage of the high spectral variability of accessory parameters like the atmosphere transmission and self-emission as well as the reflection of the environmental flux.

As a conclusion, it appears that there is no valuable reason to apply LSMWP in place of the simpler one-color pyrometry or bispectral pyrometry. All methods need *a priori* information about the emissivity. However, the requirements with one-color pyrometry (the knowledge of an emissivity level) or with bispectral pyrometry (the knowledge of the ratio of emissivity at two wavelengths) are less difficult to satisfy than the requirement with LSMWP, which is *a requirement of a strict conformity of shape* of the emissivity profile with a given parametric function, which is practically impossible to satisfy.

Regarding LSMWP, it must finally be admitted that without knowledge of the *magnitude* of emissivity, the temperature measurement cannot be very precise. Some vague intuition about the shape of the emissivity spectrum is not sufficient and to add more wavelengths does not help much. The blackbody spectrum is extremely regular; therefore, the implementation of a polynomial model for the emissivity of degree greater than 1 introduces strong correlations and generally leads to poor results.

5.3. Another multiwavelength approach: the "TES" method

The "TES" method is a multiwavelength approach that was developed for land-surface temperature evaluation through infrared remote sensing, more specifically for the Advanced Space-borne Thermal Emission and Reflection Radiometer (ASTER) on board TERRA satellite [50]. It is a five-channel multispectral thermal–IR scanner.

TES is based on the observation that the relative spectrum $\beta(\lambda) = \hat{\varepsilon}(\lambda)/\bar{\hat{\varepsilon}}$ where the apparent emissivity $\hat{\varepsilon}(\lambda)$ is obtained from an estimation of temperature $\hat{T}$ according to:

$$\hat{\varepsilon}\left(\lambda, \hat{T}\right) = \frac{L(\lambda, T) - L^{\downarrow}(\lambda)}{B\left(\lambda, \hat{T}\right) - L^{\downarrow}(\lambda)} \tag{64}$$

is relatively insensitive to the temperature estimation error. A crude estimation as with the Normalized Emissivity Method (NEM) is thus sufficient [50]. The question is then how to extract the absolute spectrum $\hat{\varepsilon}(\lambda)$ from the relative spectrum $\beta(\lambda)$. Gillespie et al. [50] found out a correlation between ε_{min} and the minimum-maximum emissivity difference defined by $MMD = \beta_{max} - \beta_{min}$:

$$\varepsilon_{min} \approx 0.994 - 0.687\, MMD^{0.737} \tag{65}$$

The regression is based on 86 laboratory reflectance spectra from the ASTER spectral library [11] for soils, rocks, vegetation, snow, and water between 10 µm and 14 µm. Ninety five percent of the samples fall within 0.02 emissivity units of the regression line. Nevertheless, this empirical relation in not universal: data related to artificial materials like metals fall far below the regression line.

After evaluating $\varepsilon_{\min}$ from the regression law, we obtain a new estimate of the emissivity spectrum from:

$$\hat{\varepsilon}(\lambda) = \beta(\lambda)\frac{\varepsilon_{min}}{\beta_{min}} \tag{66}$$

The temperature $\hat{T}$ is finally obtained by inverting Planck's law at a wavelength λ at which the emissivity profile $\hat{\varepsilon}(\lambda)$ reaches the highest value. One or two iterations are sufficient for the procedure to converge.

To be effective, TES requires at least three or four spectral bands. TES does not work well for near-grey materials (as a matter of fact ε_{min} would then stick to the value 0.994).

TES algorithm is presently used to calculate surface temperature and emissivity standard products for ASTER, which are predicted to be within respectively +1.5 K and 0.015 of correct values. Validations performed on different sites demonstrated that TES generally performs within these limits.

The regression law:

$$\varepsilon_{min} \approx 0.999 - 0.777\, MMD^{0.815} \tag{67}$$

was obtained using 108 emissivity spectra from the ASTER library, without manmade materials. It is compared with spectra of manmade materials used over urban surfaces in [57] (see Figure 11). The correlation in eq. (67) is relatively good for most considered manmade materials. Metallic surfaces are however badly modeled by this empirical relationship.

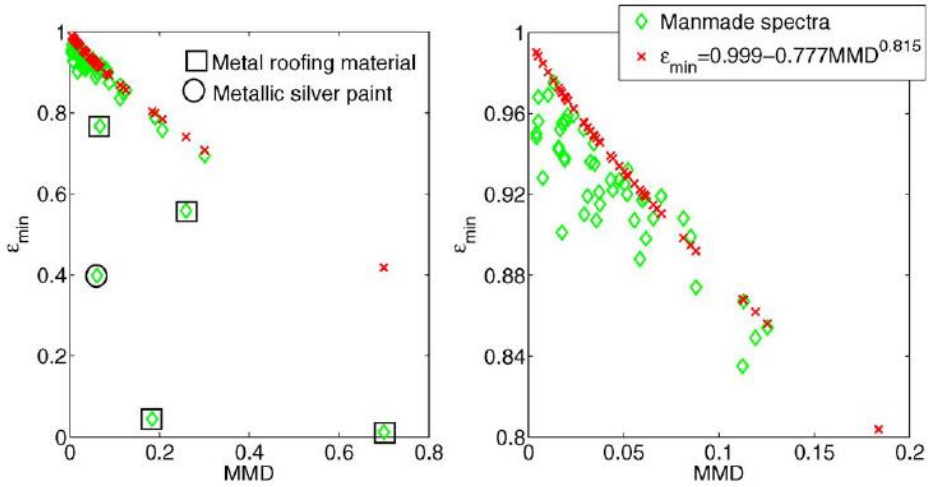


Figure 11. Correlation between MMD and ε_{min} in eq. (67) and comparison with 54 manmade materials spectra. Right figure is a detail from left plot [57].

The RMSE for emissivity is 0.017 in average (for a series of 9 manmade urban materials excluding metallic materials: brick, glass, tile, asphalt, concrete, marble, cement) and it may rise to 0.03 for some materials like marble and glass. Simultaneously, the RMSE for temperature is 0.9 K in average and may rise to 1.5 – 1.8 K for marble and glass ('true' temperature is set between 295 K and 310 K) [57].

The TES method is performing well for natural materials and manmade materials (excluding metallic materials) in the context of remote sensing. This concept could be extended to other situations. The decisive point would be to find out an empirical relation of the type shown in eq. (65) or in eq. (67) from the spectra of the considered materials.

5.4. The Bayesian approach for radiative thermometry

What has been exposed so far underlines the fact that *a priori* information on emissivity is a prerequisite for the evaluation of temperature. Actually, the Bayesian framework allows taking into account any kind of *a priori* information on the parameters to estimate, hence it should be appropriate to solve the temperature-emissivity separation problem. However, although the application of Bayesian methods to thermal characterization is relatively common today [58], it is rather rare in multispectral pyrometry and multispectral/hyperspectral infrared remote sensing [59]-[67].

In the Bayesian framework the entire problem is modeled in terms of probability in order to allow for inference, that is, instead of attempting to obtain a single solution for the interesting unknowns it offers the possibility to explore the posterior distribution to determine the uncertainty in the unknowns given the measurements and *prior* uncertainty in the unknowns. The exploration calls for computing different point estimates like the maximum *a posteriori* estimate (MAP) and the conditional mean estimate (CM) as well as marginal distributions of individual unknowns or sets of unknowns [58], [68].

The parameters $\boldsymbol{\beta}$ (vector of size $(m+1)\times 1$) and the measurements $\mathbf{Y}$ (vector of size $m\times 1$) are considered as random variables. $\pi(\boldsymbol{\beta})$ is the *prior distribution* and it represents the uncertainty of the unknown prior to obtaining the measurement. The conditional distribution of the measurements given the unknown is called the *likelihood distribution* and is denoted by $\pi(\mathbf{Y}|\boldsymbol{\beta})$. What interests us is the *posterior distribution* $\pi(\boldsymbol{\beta}|\mathbf{Y})$ which contains all information on the uncertainty of the unknowns $\boldsymbol{\beta}$ when the information on measurements $\mathbf{Y}$ is utilized [58], [68]. It is given by Bayes' theorem:

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) = \frac{\pi(\mathbf{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})}{\pi(\mathbf{Y})} \tag{68}$$

where the denominator $\pi(\mathbf{Y})$ is obtained by marginalizing $\pi(\mathbf{Y}|\boldsymbol{\beta})$ over the parameters $\boldsymbol{\beta}$. It is merely a scaling constant; since it does not involve $\boldsymbol{\beta}$ it is generally discarded for most analyses:

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \tag{69}$$

Assume that the physical model is described by eq. (18) (negligible reflection effects) and that the measurement is corrupted by additive Gaussian noise with zero mean and covariance matrix $\boldsymbol{\Omega}$, which will be noted $\pi(\boldsymbol{e}) = \mathcal{N}(0, \boldsymbol{\Omega})$ where $\boldsymbol{e}$ is the vector of the m spectral noise terms. The *likelihood distribution* $\pi(\mathbf{Y}|\boldsymbol{\beta}) = \pi(\mathbf{Y}|\boldsymbol{\varepsilon}, T)$ is then expressed by:

$$\pi(\mathbf{Y}|\boldsymbol{\varepsilon}, T) \propto exp[-(\mathbf{Y} - \boldsymbol{\varepsilon} \otimes \mathbf{B})^T \boldsymbol{\Omega}^{-1} (\mathbf{Y} - \boldsymbol{\varepsilon} \otimes \mathbf{B})/2] \tag{70}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1 \ \dots \ \varepsilon_m)^T$, $\mathbf{B} = \big(B(\lambda_1, T) \ \dots \ B(\lambda_m, T)\big)^T$ and $\otimes$ denotes the elementwise product.

On the other side, regarding the prior, we will assume that emissivity and temperature are independent variables, hence $\pi(\boldsymbol{\beta}) = \pi(\boldsymbol{\varepsilon})\pi(T)$. For ease, we consider that the spectral emissivities are independent as well.

5.4.1. A simple example: single-color pyrometry

When there is only one spectral measurement, the *posterior distribution* $\pi(\boldsymbol{\beta}|\mathbf{Y})$ reduces to:

$$\pi(\varepsilon, T|Y) \propto exp\left[-\big(Y - \varepsilon B(\lambda, T)\big)^2 / 2\sigma^2\right] \pi(\varepsilon)\pi(T) \tag{71}$$

where σ is the noise RMS. If we are only interested in temperature, emissivity is then considered as a nuisance parameter. To obtain the posterior distribution for temperature alone, we thus have to marginalize the joint distribution $\pi(\varepsilon, T|Y)$ with respect to emissivity.

Let us consider for ease a uniform *a priori* distribution for emissivity: $\pi(\varepsilon) = \mathcal{U}(\varepsilon_{min}, \varepsilon_{max})$. The marginal posterior distribution related to temperature, $\pi(T|Y)$, can thus be expressed analytically [59]:

$$\pi(T|Y) \propto \frac{\pi(T)}{B(\lambda,T)}\left[erf\left(\frac{Y-\varepsilon_{max}B(\lambda,T)}{\sqrt{2}\sigma}\right) - erf\left(\frac{Y-\varepsilon_{min}B(\lambda,T)}{\sqrt{2}\sigma}\right)\right] \tag{72}$$

As an example, let us consider a monochromatic sensor at 4.7 µm with 3 % RMS noise, and assume that emissivity is expected to be in the range [0.5, 0.75]. If the measured radiance corresponds to the radiance emitted by a surface at 600 K with an emissivity of 0.6, the posterior distribution for temperature, given that measurement, is described by the black curve in Figure 11 (a non-informative prior is considered for temperature, namely $\pi(T) = \mathcal{U}(500\ K, 700\ K)$). The curve is quite asymmetrical; the maximum a posteriori estimate is 582 K whereas the conditional mean estimate is 597 K, which is closer to the real value 600 K. The distribution is quite large since the *a priori* distribution of the emissivity is large itself and the *a priori* distribution of temperature is non-informative. To shrink the *a posteriori* distribution we should have better information on the emissivity and possibly on temperature.

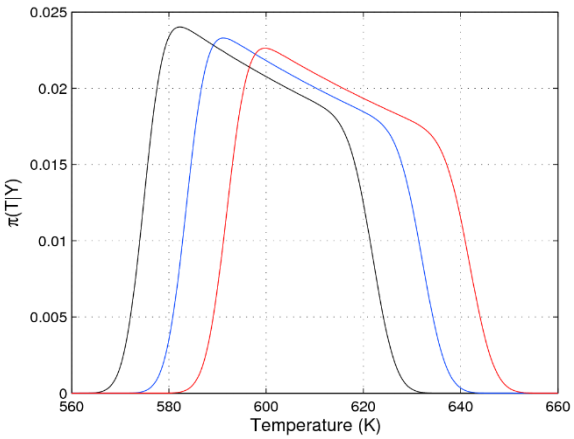


Figure 11. Posterior distribution of temperature in the case of uniform prior distribution of temperature and emissivity: $\pi(\varepsilon) = \mathcal{U}(0.5{,}0.75)$. The measured radiance corresponds to the emitted radiance of a surface at 600 K with an emissivity of 0.6 (black), 0.65 (blue), and 0.7 (red). Noise RMS is 3 %.

Let us now consider a measured radiance that is 8.3 % higher than before. Among the infinite number of possibilities, it could correspond to the radiance emitted by a surface at 600 K with an emissivity of 0.65. The posterior distribution for temperature is now given by the blue curve in Figure 11. The maximum *a posteriori* has risen to 592 K and the conditional mean estimate to 606 K. Let us pursue the analysis. If the measured radiance was 16.7 % higher than the initial value (it could now correspond to the radiance emitted by a surface at 600 K with an emissivity of 0.7), the posterior distribution for temperature is then given by the red curve. The

maximum *a posteriori* has risen further to 600 K and the conditional mean estimate to 615 K. Notice that for a progressively higher RMS noise, the curves would be progressively more rounded and approach a Gaussian curve.

As a final remark, let us say that thanks to the availability of *a priori* information on emissivity or temperature, the Bayesian approach allows to "anchor" the solution instead of providing an infinite set of equally acceptable solutions (ε, T). Nevertheless the "anchoring" is not bound to a particular solution set (ε, T), but rather loose. The poorer the *a priori* information, the more the "anchoring" is loose.

5.4.2. Multiwavelength pyrometry (linear approximation)

The implementation of the linear approximation for multiwavelength pyrometry (notably by introducing Wien's law and considering the logarithm of the spectral signals S_i, $i = 1, \ldots, m$ as the observables Y_i, $i = 1, \ldots, m$, see eq. (35)) has the advantage, considering normal distributions for the priors and for the measurement noise, to yield analytical expressions.

Equation (35) can be rewritten as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}' \tag{73}$$

where $\boldsymbol{\beta} = \left[ln(\varepsilon_1) \ \ldots \ ln(\varepsilon_m) \ \ T_{ref}/T\right]^T$ is the vector of (linear) parameters whose prior is a multivariate normal distribution of covariance matrix $\mathbf{W}$, namely $\pi(\boldsymbol{\beta}) = \mathcal{N}\left(\boldsymbol{\beta}^{prior}, \mathbf{W}\right)$, whereas $\boldsymbol{e'}$ is the vector of additive errors with $\pi(\boldsymbol{e'}) = \mathcal{N}(0, \boldsymbol{\Omega})$ and $\mathbf{X}$ is the sensitivity matrix:

$$\mathbf{X} = (\mathbf{I}_{mm} \quad -\boldsymbol{\mu}_{\mathrm{m1}}) \ ; \quad \boldsymbol{\mu} = (\mu_1 \ \mu_2 \ \ldots \mu_m)^T \tag{74}$$

where the constant coefficients μ_i, $i = 1, \ldots, m$ have been defined in eq. (36).

The posterior distribution $\pi(\boldsymbol{\beta}|\mathbf{Y})$ in eq. (69) becomes:

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) \propto exp\left(-\frac{1}{2}\left((\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{prior}\right)^T\mathbf{W}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{prior}\right)\right)\right) \tag{75}$$

In the linear Gaussian case, all conditional distributions are Gaussian. It suffices therefore to compute the (conditional) means and covariances only (the maximum a posteriori estimator is equal to the (conditional) mean estimator) [68]. To obtain the maximum *a posteriori* (MAP) estimator we differentiate the argument of the exponential in eq. (75) with respect to the parameter vector and look for the parameter vector that makes it vanish [68]. In the end we have:

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) \propto exp\left(-\frac{1}{2}\left(\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{MAP}\right)^T{\Gamma_{\boldsymbol{\beta}|\mathbf{Y}}}^{-1}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{MAP}\right)\right)\right) \tag{76}$$

with the following expression for the MAP estimator:

$$\widehat{\boldsymbol{\beta}}_{MAP} = \Gamma_{\boldsymbol{\beta}|\mathbf{Y}}\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{Y} + \mathbf{W}^{-1}\boldsymbol{\beta}^{prior}\right) \tag{77}$$

where $\Gamma_{\boldsymbol{\beta}|\mathbf{Y}}$ is the posterior covariance matrix:

$$\Gamma_{\boldsymbol{\beta}|\mathbf{Y}} = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X} + \mathbf{W}^{-1})^{-1} \tag{78}$$

Alternative expressions are [68]:

$$\widehat{\boldsymbol{\beta}}_{MAP} = \boldsymbol{\beta}^{prior} + \mathbf{W}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{W}\mathbf{X}^T + \boldsymbol{\Omega})^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{prior}\right) \tag{79}$$

and:

$$\Gamma_{\boldsymbol{\beta}|\mathbf{Y}} = \mathbf{W} - \mathbf{W}\mathbf{X}^T(\mathbf{X}\mathbf{W}\mathbf{X}^T + \boldsymbol{\Omega})^{-1}\mathbf{X}\mathbf{W} \tag{80}$$

Notice that since $\mathbf{X}$ is full row rank rectangular matrix, and knowing that $\boldsymbol{\Omega}$ is positive definite, $\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}$ is a singular (not invertible) matrix. However, adding the positive definite matrix $\mathbf{W}^{-1}$ makes the matrix $(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X} + \mathbf{W}^{-1})$ invertible. Hence, the *prior information* provides the *regularization* needed since $\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}$ is singular. Beyond the presently underdetermined problem, it provides the regularization needed when $\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}$ is ill-conditioned.

Eqs. (77) and (78) show that when the prior variance decreases while the measurement error variance is kept constant, the *a priori* solution progressively dominates the solution.

The following example intends to illustrate that multispectral measurements can lead to valuable results when combined to priors of good quality, at least for some of them.

The simplest case of bicolor pyrometry has been considered. The measurements are assumed to be performed at 3.7 µm and 4.7 µm with a noise RMS of 5 %. The emissivity priors have mean values of 0.75 and 0.45 at the first and second wavelength, respectively. In both cases the standard deviation is 0.1. On the other side, the temperature prior has a mean value of 650 K and a quite large standard deviation, namely 150 K in order to express that the temperature is not well-known beforehand.

Consider now that the two measured spectral signals correspond to the radiances emitted by a surface at 600 K with spectral emissivities of 0.7 and 0.5 (this will be referred as the set of "true" – unknown – values, which, we hope, the estimators come close to). What are the MAP estimators and the uncertainty for temperature and emissivity? The application of the eqs. (77)-(80) provides the answer which is summarized in Table 3.

Table 3. Results of estimation in the case of bicolor pyrometry.

Parameter	MAP estimator	stand. deviation
Temperature	598 K	11.7 K
emissivity at 3.7 µm	0.72	0.09
emissivity at 4.7 µm	0.51	0.05

Figure 12 illustrates the marginal distributions for the three parameters (*a priori* and *a posteriori*). Notice that normal distributions apply to the transformed parameters $\left[ln(\varepsilon_1) \ \dots \ ln(\varepsilon_m) \ \ T_{ref}/T\right]^T$; a backward transformation has been performed to plot the distributions of $(\varepsilon_1 \ \dots \ \ \varepsilon_m \ \ T)^T$.

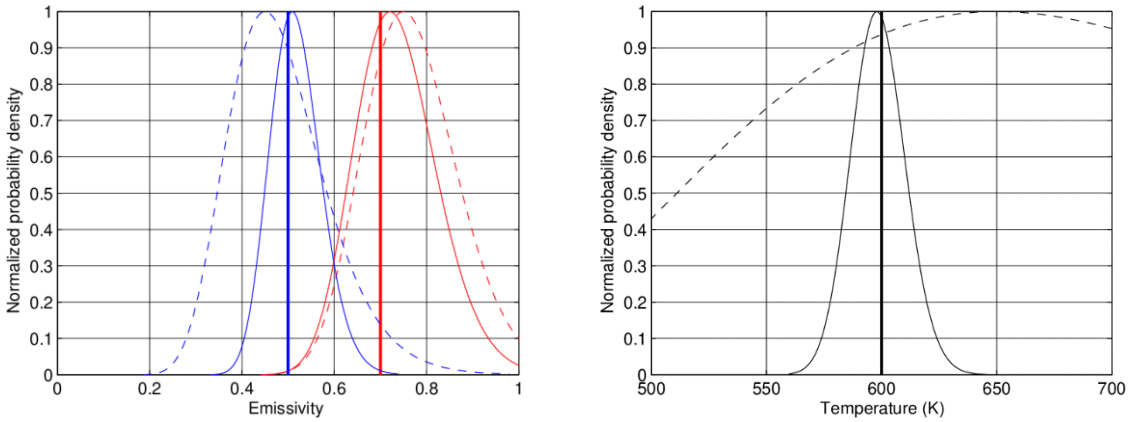


Figure 11. Left: normalized probability density of the emissivity at 3.7 µm (in red) and at 4.7 µm (in blue). The prior density is in dashed line, the posterior density is in continuous line. The "true" (unknown) values of the two emissivities (resp. 0.7 and 0.5) are indicated by vertical bold lines. Right: normalized probability density of the temperature. The prior density is in dashed line, the posterior density is in continuous line. The "true" value of temperature (600 K) is indicated by a vertical bold line.

Despite a temperature prior of poor quality, the estimators are quite close to the "true" values. Furthermore, if we compare the *a priori* distributions and the *a posteriori* distributions, we notice that the move is indeed towards the "true" values. In addition, the measurements contribute to shrink the distributions (all variances have decreased).

The simple analysis that has been performed so far can be extended to more than two wavelengths without any difficulty. The signals from multispectral or hyperspectral detectors can thus be processed and inverted directly (*i.e.* without iterations) through simple matrix algebra.

By the way, the former example clearly showed that *a prior* information on the magnitude of the spectral emissivities is of much higher value than *a prior* information that the emissivity profile belongs to a particular class of profiles (*e.g.* polynomial functions).

5.4.3. Multiwavelength radiometry (non-linear case)

The linear case developed in §5.4.2 presents a few limitations. One could argue that Wien's law is a mere approximation of (exact) Planck's law, however, as mentioned in §2.1, the approximation error is very small as long as the product λT is less than about 3 000 µm·K, which is the case when the spectral range is chosen in the rising part of the blackbody-radiance curve, namely where the sensitivity of the radiance to temperature is highest. The limitations

come rather from the fact that in reality the emissivity and temperature priors are not necessarily Gaussians. As a matter of fact, some of the "theoretical" distributions in Figure 11-left go beyond the boundary ε =1, which is unrealistic. For the emissivity, truncated prior distributions should thus be implemented. Moreover, we should be able to simulate probability distributions of arbitrary shape.

Beyond linear and Gaussian models, Bayesian inference requires a statistical estimation of the posterior probability distributions which involves numerical sampling. Markov Chain Monte Carlo (MCMC) algorithms are implemented for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. Metropolis–Hastings algorithm [60] and Gibbs' sampler [62], [64], [65] are examples of MCMC algorithms.

The versatility of the MCMC algorithms makes them capable of handling radiative problems more complex than those described by the simple "pyrometric" equation in eq. (18). As such, the reflection contribution could be added in the unknown parameters since a good prior is generally accessible.

More details on MCMC algorithms can be found in the lecture devoted to Bayesian inference.

6. Conclusion

Accurate temperature measurement by radiative means is not an easy task. Many parameters have to be evaluated beforehand to extract the surface emitted radiance from the measured radiance (atmospheric contributions: self-emission and attenuation, reflections from the environment). We then face the problem of temperature-emissivity separation. This underdetermined problem requires that some knowledge about the emissivity of the tested material is introduced. The general feeling is that multiplying the spectral measurements at different wavelengths would help identify the temperature. The underdetermined nature of the problem is however invariably maintained. Introducing a model of the emissivity spectral profile is often a misleading idea: high systematic errors inevitably occur when the model does not correspond perfectly to the real emissivity profile. Having some knowledge about the *magnitude* of emissivity is much more useful (but unfortunately more demanding) than imposing a particular class of *shapes*. The Bayesian framework is definitely well-suited to this task.

7. References

[1] Siegel R and Howell J R 1972 *Thermal Radiation Heat Transfer* (McGraw Hill).

[2] Chrzanowski K and Szulim M 1998 Measure of the influence of detector noise on temperature measurement accuracy for multiband infrared systems *Applied Optics* **37**(22) 5051-5057

[3] Krapez J.-C. 2011 Radiative measurements of temperature in *Thermal Measurements and Inverse Techniques* (Taylor & Francis)

[4] Berk A., Conforti P., Kennett R., Perkins T., Hawes F., Van Den Bosch J. 2014 MODTRAN® 6: A major upgrade of the MODTRAN® radiative transfer code. *6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing WHISPERS* pp. 1-4

[5] Labarre L., Caillault K., Fauqueux S., Malherbe C., Roblin A., Rosier B., Simoneau P. 2010 An overview of MATISSE-v2.0 *Optics in Atmospheric Propagation and Adaptive Systems XIII* SPIE Vol. 7828, p. 782802

[6] Loarer T., Greffet J.-J. and Huetz-Aubert, M. 1990 Noncontact surface temperature measurement by means of a modulated photothermal effect, *Appl. Optics* **29**(7) 979-987

[7] Loarer T., Greffet J.-J. 1992 Application of the pulsed photothermal effect to fast surface temperature measurements, *Appl. Optics* **31**(25) 5350-5358

[8] Amiel S., Loarer T., Pocheau C., Roche H., Aumeunier M. H., Gauthier E., LeNiliot C. and Rigollet, F. 2012 Surface temperature measurement of plasma facing components with active pyrometry. In Journal of Physics: Conference Series (Vol. 395, No. 1, p. 012074). IOP Publishing

[9] Amiel S., Loarer T., Pocheau C., Roche H., Gauthier E., Aumeunier M. H., Le Niliot C. Rigollet F., Courtois X., Jouve M., Balorin C. and V. Moncada V. 2014 2D surface temperature measurement of plasma facing components with modulated active pyrometry. *Rev. Sci. Instr.,* 85(10), 104905

[10] Touloukian Y S and DeWitt D P 1970 *Thermal radiative properties. Thermophysical properties of matter* (Plenum Corp. New-York)

[11] Salisbury J W and d'Aria D M 1992 Emissivity of terrestrial materials in the 8-14 µm atmospheric window *Remote Sens. Environ.* **42** 83-106

[12] Baldridge A M, Hook S J, Grove C I and Rivera G 2009 The ASTER Spectral Library Version 2.0 *Remote Sens. Environ.* **113** 711-715 http://speclib.jpl.nasa.gov/

[13] Corwin R R and Rodenburgh A 1994 Temperature error in radiation thermometry caused by emissivity and reflectance measurement error *Applied Optics* **33**(10) 1950-1957

[14] Hervé P and Sadou A 2008 Determination of the complex index of refractory metals at high temperatures: application to the determination of thermo-optical properties *Infrared Physics & Technology* **51** 249–255

[15] Pierre T, Rémy B and Degiovanni A 2008 Microscale temperature measurement by the multispecrtral and statistic method in the ultraviolet-visible wavelengths *J. Appl. Phys.* **103** 034904-1-10

[16] Duvaut T, Georgeault D and Beaudoin J L 1996 Pyromètre multispectral infrarouge : application aux métaux *Rev. Gen. Therm* **35** 185-196

[17] Hernandez D, Sans JL, Netchaieff A, Ridoux P, Le Sant V 2009 Experimental validation of a pyroreflectometric method to determine the true temperature on opaque surface without hampering reflections *Measurement* **42** 836-843

[18] Sentenac T., Gilblas R., Hernandez D., Le Maoult Y. 2012 Bi-color near infrared thermoreflectometry: A method for true temperature field measurement *Rev. Sci. Instr.* 83(12), 124902

[19] Sentenac T., Gilblas R., & Bugarin F. 2019 Trichromatic thermoreflectometry for an improved accuracy of true temperature field measurement on a multi-material part. *Int. J. Therm. Sci.* 145, 105980

[20] Krapez J-C, Bélanger C and Cielo P 1990 A double-wedge reflector for emissivity enhanced pyrometry *Meas. Sci. Technol* **1** 857-864

[21] Foley G.M. 1978 *High Temp. High Press*. **10**:391

[22] Watari M. Watanabe Y., Chigira S., Tamura Y., *Yokogawa Tech. Rep.* **29**:25

[23] Anderson 1985 *Adv. Instrument.* **40**:1337

[24] Tsai B.K., Shoemaker R.L., DeWitt D.P., Cowans B.A., Dardas Z., Delgass W.N., Dail G.J. 1990 Dual Wavelength radiation thermometry: emissivity compensation algorithms, *Int. J. Thermophysics* **11**(1) 269-281

[25] Gardner J L 1980 Computer modelling of a multiwavelength pyrometer for measuring true surface temperature *High Temp – High Press*. **12** 699-705

[26] Coates P B 1981 Multiwavelength pyrometry *Metrologia* **17** 103-109

[27] Gardner J L, Jones T P, Davies M R 1981 A six wavelength pyrometer, *High Temp – High Press*. **13** 459-466

[28] Hunter B, Allemand C D and Eager T W 1985 Multiwavelength pyrometry: an improved method *Opt. Eng.* **24**(6) 1081-1085

[29] Hunter B, Allemand C D and Eager T W 1986 Prototype device for multiwavelength pyrometry *Opt. Eng.* **25**(11) 1222-1231

[30] Hiernault J P, Beukers R, Heinz W, Selfslag R, Hoch M and Ohse R.W. 1986 Submillisecond six-wavelength pyrometer for high temperature measurements in the range 2000K-5000K *High Temp – High Press.* **18** 617-625

[31] Nordine P C 1986 The accuracy of multicolour optical pyrometry *High Temp. Sci.* **21** 97-109

[32] DeWitt D P and Rondeau R.E. 1989 Measurement of surface temperatures and spectral emissivities during laser irradiation *J. Thermophysics* **3**(2) 153-159

[33] Tank V and Dietl H 1990 Multispectral infrared pyrometer for temperature measurement with automatic correction of the influence of emissivity *Infrared Physics* **30**(4) 331-342

[34] Khan MA, Allemand C and Eager TW 1991 Noncontact temperature measurement. I. Interpolation based techniques *Rev. Sci. Instrum* **62**(2) 392-402

[35] Khan MA, Allemand C and Eager TW 1991 Noncontact temperature measurement. II. Least square based techniques Rev. Sci. Instrum **62**(2) 403-409

[36] Gathers G.R. 1991 Analysis of multiwavelength pyrometry using nonlinear least square fits and Monte Carlo methods *11th Symp. Thermophysic. Prop.* Boulder June 1991

[37] Lindermeir E, Tank V and Hashberger P 1992 Contactless measurement of the spectral emissivity and temperature of surfaces with a Fourier transform infrared spectrometer *Proc. SPIE* 1682 354-364

[38] Duvaut T, Georgeault D and Beaudoin JL 1995 Multiwavelength infrared pyrometry: optimization and computer simulations *Infrared Phys.& Technol*. **36** 1089-1103

[39] Chrzanowski K and Szulim M 1998 Error on temperature measurement with multiband infrared systems *Applied Optics* **38**(10) 1998-2006

[40] Chrzanowski K and Szulim M 1999 Comparison of temperature resolution of single-band, dual-band and multiband infrared systems *Applied Optics* **38**(13) 2820-2823

[41] Scharf V, Naftali N, Eyal O, Lipson S G and Katzir A 2001 Theoretical evaluation of a four-band fiber-optic radiometer *Appl Opt.* **40**(1) 104-111

[42] Mazikowski A and Chrzanowski K 2003 Non-contact multiband method for emissivity measurement *Infrared Phys. & Technol.* **44** 91-99

[43] Cassady L D and Choueiri E Y 2003 High Accuracy Multi-color Pyrometry for High Temperature Surfaces *IEPC-03-79 28th Int. Electric Propulsion Conference* Toulouse France March 17-21 2003

[44] Wen C D and Mudawar I 2004 Emissivity characteristics of roughened aluminium alloy surfaces and assessment of multispectral radiation thermometry (MRT) emissivity models *Int. J. Heat Mass Transfer* **47** 3591-3605

[45] Wen C D and Mudawar I 2004 Emissivity characteristics of polished aluminium alloy surfaces and assessment of multispectral radiation thermometry (MRT) emissivity models *Int. J. Heat Mass Transfer* **48** 1316-1329

[46] Sade S and Katzir A 2004 Multiband fiber optic radiometry for measuring the temperature and emissivity of gray bodies of low or high emissivity *Appl. Opt.* **43**(9) 1799-1810

[47] Uman I and Katzir A 2006 Fiber-optic multiband radiometer for online measurements of near room temperature and emissivity *Optic Letters* **31**(3) 326-328

[48] Duvaut T 2008 Comparison between multiwavelength infrared and visible pyrometry: application to metals *Infrared Phys.& Technol*. **51** 292-299

[49] Rodiet C., Remy B., Pierre T., & Degiovanni A. 2015 Influence of measurement noise and number of wavelengths on the temperature measurement of opaque surface with variable emissivity by a multi-spectral method based on the flux ratio in the infrared-ultraviolet range. *High Temp.--High Press.* **44**(3)

[50] Rodiet C., Remy B., & Degiovanni A. 2016 Optimal wavelengths obtained from laws analogous to the Wien's law for monospectral and bispectral methods, and general methodology for multispectral temperature measurements taking into account global transfer function including non-uniform emissivity of surfaces. *Infrared Phys.& Technol*. **76** 444-454

[51] Daniel K., Feng C., Gao, S. 2016 Application of multispectral radiation thermometry in temperature measurement of thermal barrier coated surfaces. *Measurement*, **92**, 218-223

[52] Zhang C., Gauthier E., Pocheau C., Balorin C., Pascal J. Y., Jouve M., Aumeunier M.H., Courtois X., Loarer T., Houry M. 2017 Surface temperature measurement of the plasma facing components with the multi-spectral infrared thermography diagnostics in tokamaks. *Infrared Phys.& Technol*. **81**, 215-222

[53] Bouvry B., Cheymol G., Ramiandrisoa L., Javaudin B., Gallou C., Maskrot H., Horny N., Duvaut T., Destouches C., Ferry L., Gonnier C. 2017 Multispectral pyrometry for surface temperature measurement of oxidized Zircaloy claddings. *Infrared Phys.& Technol*. **83**, 78-87

[54] Barducci A and Pippi I 1996 Temperature and emissivity retrieval from remotely sensed images using the "grey body emissivity" method *IEEE Trans. Geosci. & Remote Sensing* **34**(3) 681-695

[55] Gillespie A., Rokugawa S., Matsunaga T., Cothern J.S., Hook S., Kahle A.B. 1998 A temperature and emissivity separation algorithm for advanced spaceborne thermal

emission and reflection radiometer (ASTER) images *IEEE Trans. Geosci. Remote Sens.* **36**(4): 1113-1126

[56] Sabol D.E., Gillespie A.R., Abbott E., Yamada G. 2009 Field validation of the ASTER temperature and emissivity separation algorithm *Remote Sens. Environ*. **113**: 2328-2344

[57] Oltra-Carrio R., Cubero-Castan M., Briottet X., Sobrino J. 2014 Analysis of the performance of the TES algorithm over urban areas *IEEE Trans. Geosci. Remote Sens.* **52**(11) 6989-6998

[58] Kaipio J. P., Fox C. 2011 The Bayesian framework for inverse problems in heat transfer *Heat Transf. Eng.* **32**(9), 718-753

[59] Morgan J. A. 2005 Bayesian estimation for land surface temperature retrieval: The nuisance of emissivities. *IEEE Trans. Geosci. Remote Sens***. 43**(6) 1279-1288

[60] Heasler P., Posse C., Hylden J., Anderson K. 2007 Nonlinear bayesian algorithms for gas plume detection and estimation from hyper-spectral thermal image data. *Sensors*, **7**(6), 905-920

[61] Morgan J. A. 2011 Comparison of Bayesian land surface temperature algorithm performance with Terra MODIS observations *Int. J. Remote Sens.* **32**(23) 8139-8159

[62] Berrett C., Williams G. P., Moon T., & Gunther J. 2014 A Bayesian Nonparametric Model for Temperature-Emissivity Separation of Long-Wave Hyperspectral Images *Technometrics* **56**(2) 200-211

[63] Krapez J.-C. 2015 Measurements without contact in heat transfer: principles, implementation and pitfalls *METTI 6* Advanced School:Thermal Measurements and Inverse Techniques, Biarritz, March 1- 5, 2015

[64] Ash J. N., Meola J. 2016 Temperature-emissivity separation for LWIR sensing using MCMC. *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXII SPIE* Vol. 9840, p. 98401O

[65] Toullier T., Dumoulin J., & Mevel L. 2018 Etude de sensibilité de différentes méthodes de séparation pour l'évaluation simultanée de l'émissivité et de la température par thermographie infrarouge multispectrale, Congrès SFT, Pau.

[66] Pierre T., Krapez J.-C., Orlande H. R. B., Rodiet C., Le Maux D., Courtois M., Le Masson, P, Lamien B., Simultaneous Estimation of Temperature and Emissivity of Metals around Their Melting Points by Deterministic and Bayesian Techniques, *International Journal of Heat and Mass Transfer*, vol. 183, février 2022, p. 122077.

[67] Pierre T., Krapez J.-C., Orlande H. R. B., Rodiet C., Le Maux D., Courtois M., Le Masson, P, Lamien B., Multiple inversion techniques with multispectral pyrometry for the estimation of temperature and emissivity of liquid niobium and 100c6 steel, *Heat Transfer Engineering*, accepted for publication.

[68] Kaipio J., Somersalo E. 2006 *Statistical and computational inverse problems* (Vol. 160). Springer Science & Business Media.

Lecture 4 Part B: Quantitative Infrared Thermography

H. Pron[1], L. Ibos[2]

1. ITheMM, Université de Reims, Reims, France
2. CERTES, IUT de Sénart-Fontainebleau, Université Paris-Est Créteil, Moissy-Cramayel, France

E-mail: herve.pron@univ-reims.fr
ibos@u-pec.fr

Abstract. The main objective of this lecture is to make the end users aware of the various physical phenomena and especially of the errors frequently met during temperature and heat flow measurement by infrared thermography. For that purpose, this chapter will present the three aspects of a quantitative infrared measurement that are spatial, temporal and thermal resolution. First, the spatial resolution will be discussed, showing that an increase of the matrix size does not necessarily induce an improvement of the spatial resolution. Then, a paragraph is especially dedicated to the temporal aspects, as far as many applications require at least stable frequency to high speed imaging. Last but not least, the calibration of the systems is discussed, showing that accurate measurements often need a specific home-made thermal calibration.

List of acronyms:

- **SRF**: Slit Response Function
- **IR**: Infrared
- **ADC**: Analog-Digital Conversion
- **BPR**: Bad Pixel Replacement
- **NUC**: Non-Uniformity Correction
- **CNUC** Compensated Non-Uniformity Correction

Scope

1. Foreword: why it is important to well know your equipment?

Prior to any quantitative measurement using an infrared device, it is important to be aware of the limitations of the technique, but also of the transfer function of the device. Some work, concerning either the thermography technique [1,2] or the associated metrology [3-6] are available in the literature.
There are three major points of necessary characterization of the devices: inaccuracies of the calibration, spatial non-uniformity, and irregular time sampling can lead to false parameter estimation.

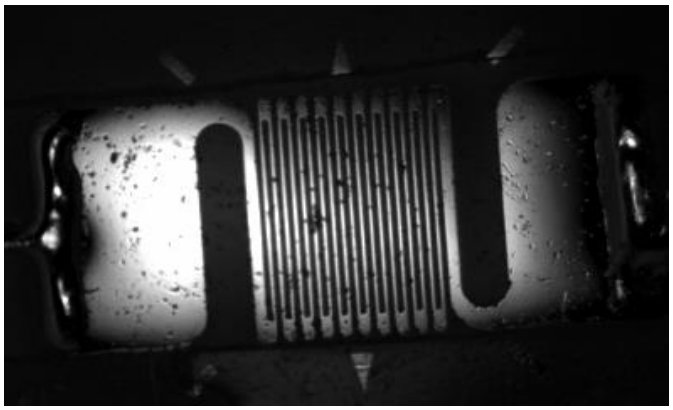


Figure 1 : *Strain gauge (tracks of approximately 20 µm) observed with a "M1" lens*.

2. Spatial resolution

The focal plane array technology has indubitably led to improvements in image quality (Figure 1 hereafter). However, the quality of an image can be considered either from the point of view of the aesthetic, or from the one of the metrology. Unfortunately, these two approaches are rarely compatible…

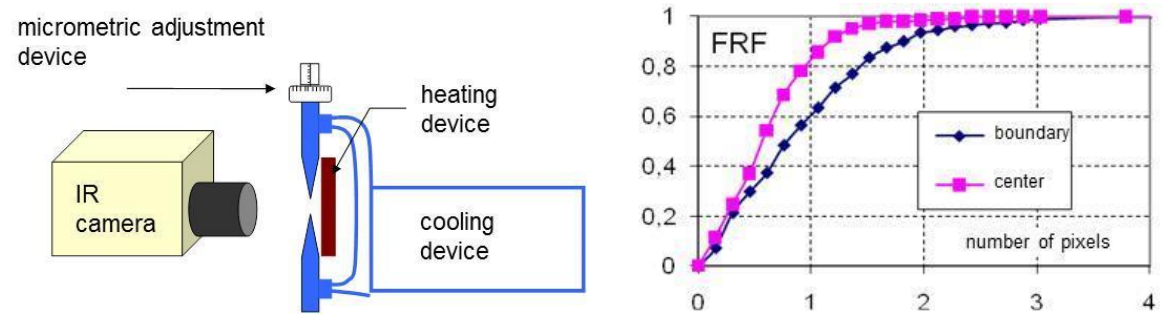
In order to ensure to obtain reliable measurements, the independence of each sensor relatively to its neighbors must be checked. One of the most current tests for characterizing such equipment is the Slit Response Function (SRF) test: the camera focuses on a thermal side-cooled slit of variable width, placed in front of a hot plate; the following contrast function is then studied:

$$SRF = \frac{V(x) - V_{\min}}{V_{\max} - V_{\min}}, \tag{1}$$

where *V(x)* is the value recorded for a slit width equal to *x*, V_{max} is the value recorded when the slit is wide open ($x \to \infty$) and V_{min} is the recorded value on the cooled part (Figure 2a). In general, it is assumed that, for 320 x 240 pixel cameras, to obtain a good measurement the object must be projected on at least two detectors. Thus, with a lens magnification of 1 ("M1") and a matrix periodicity of 30 µm, one obtains truly independent information only at each step of 60 µm.

A study of this SRF for different positions clearly shows (Figure 2) that the pixels are quite more correlated on the edges of the array than in the center. Note that there is indeed a

problem of correlation between close measurement points, i.e. on the one hand, only the contrast (and by no means the average value) is affected and, on the other hand, there is convolution of the thermal scene by this response function. Consequently, a simple geometrical correction (e.g. of repositioning of the points in the image, or amplification and/or offsets applied to each pixel) is necessary to recover the real quantitative image of the scene, in addition to a deconvolution procedure. A possible restoration procedure of thermal images based on the characterization of the Modulation Transfer Function of the camera was proposed in [7].



(a) *(b)*

Figure 2 : *(a) Slit Response Function; (b) SRF near the edge of the array compared to SRF at the centre (CEDIP IRC 320-4 LW camera)*

3. Temporal analysis

First, it seems to be necessary to remind, before any characterization, some important definitions concerning this technology.

- Integration time: this duration corresponds to the part of the image period during which the detectors are effectively loading their associated capacitors; so, they measure the external infrared radiation during the integration time only. Any user should be aware that this duration is very short (often about one millisecond), compared to the frame period (typically about 10 or 20 ms for a full window): most of the frame period is dedicated to the reading of the stored electrical signals. In some situations, this is a problem since very quick phenomena can occur during this "blind" phase.

- Multiplexing duration: the reading of the different pixels is not simultaneous; according to the considered device, the pixel signals can be read by up to four channels, at a sampling rate of a few MHz (f_{ADC} hereafter). Then, t_s being the settle time, the maximum frame frequency can be simply obtained by:

$$f = \left(t_i + \frac{n_r \times n_c}{N \times f_{ADC}} + t_s \right)^{-1} \quad (2)$$

where:
n_r is the number of rows,
n_c the number of columns,
and N the number of channels.

The integration time t_i is the duration during which the radiation coming from the thermal scene is collected by the detectors of the camera. Consequently, it determines the ultimate temporal resolution of the device. As the image transfer time to the storage memory or to the hard disk is often much higher than the integration time (several milliseconds compared to some tenths or hundreds of microseconds), the detector thus does not see the scene during most of the time, which is particularly penalizing for observing fast phenomena.

Regardless of the problems connected to the integration time, the temporal analysis can be disturbed by the absence of some images in the stored sequence. Depending on the devices, a temporal shift of one or two images can occur at the beginning of the sequence. This is due to the fact that the first stored image corresponds to the one that was captured when the starting order occurred, not the actual image at the beginning of the sequence; sometimes, due to pre-processing, the temporal shit can be of two images. Then, on condition that the user is aware of this fact, a simple sequence shift is enough to correct this edge effect.

The second, more penalizing, problem is the absence of some images within a sequence. This problem is relatively unimportant in terms of visualization, but can become critical in the data processing when time is highly involved. Algorithms that are compatible with variable acquisition frequencies are then required. To count and isolate times from the missing images, it is possible to directly read time information in the files from the camera, provided that they have been accurately stored, *i.e.* sufficient with respect to the acquisition frequencies used. Depending on the camera model, the number of images missing can thus range from one to several dozens.

Figure 3 presents the artefacts observed in the case of a numerical lock-in procedure applied to a series of 30 images in which only two images are missing. If the amplitude is not very affected, the phase has a completely erratic behaviour, and takes a value that depends directly on the number and the phase of the missing images.

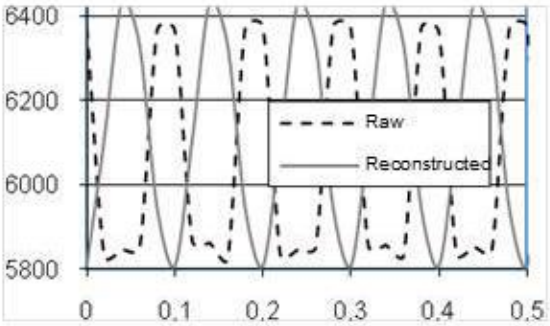


Figure 3 : *Errors induced by the missing images in lock-in thermography: amplitude is not really affected but phase is strongly distorted*

4. Thermal aspects

4.1. Thermal noise and thermal drift

The infrared devices usually used in R&D are cooled at approximately 80 K in order to reduce radiation in the vicinity of the infrared sensors. In new-generation IR cameras, a Stirling cycle

engine has replaced liquid nitrogen cooling systems of older cameras. Though the cameras have thus gained in portability, this new system has a non-negligible drawback: the cooling, which was quasi-instantaneous with nitrogen, now requires at least 10 min before any measurement is possible (figure 4a).

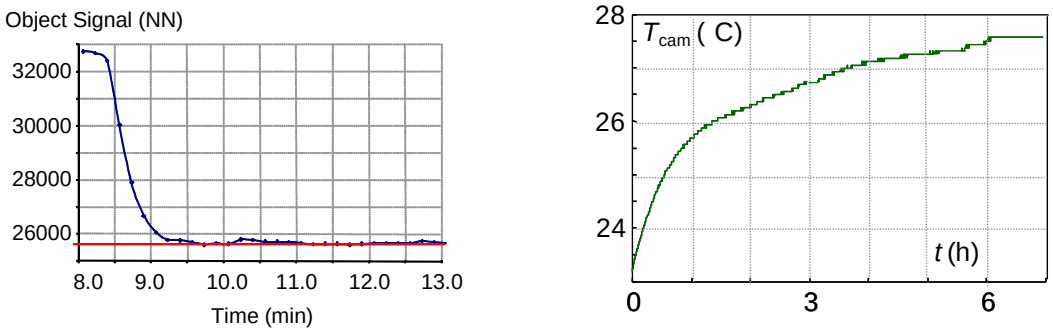


Figure 4 : *(a) Cooling CEDIP IRC320-4LW, (b) thermal drift CEDIP JADE III*

In addition, once the cooling is achieved, a slow drift of about 1 to 5 mK per second can occur with certain materials, sometimes over durations reaching a few hours (figure 4b). This temperature drift is mainly due to the evolution of the internal temperature of the camera, and modify the sensor responses, so it is appropriate in several situations to wait until the camera temperature is stabilized, or to take this internal drift into account in the conversion of the digitized signal into temperature (Compensated NUC). In addition, certain lower quality materials have instabilities of 0.5 or even 1 K, which that is incompatible with quantitative measurements.

4.2. Environment thermal stability

The signal measured by a camera comes primarily from the object (assumed to be gray and opaque in the camera's spectral range), but also, to a lesser extent (in the most favorable conditions), from the environment and atmosphere (figure 5). If the environment can be considered as an integral radiator of temperature T_{env} and if the atmosphere between the target and the camera is isothermal at the temperature T_{atm}, considering a coefficient of transmission τ_{atm}, the measured intensity L_{mes} can be formulated as a function of the intensity L^0 of a blackbody at the object temperature:

$$L_{\text{mes}} = \tau_{\text{a}} .\varepsilon .L^0(T_{\text{obj}}) + \tau_{\text{a}}(1-\varepsilon)L^0(T_{\text{env}}) + (1-\tau_{\text{a}})L^0(T_{\text{atm}}) \tag{3}$$

For short distance measurements (about a few tenths of cm), the atmosphere can reasonably be considered as being transparent, and thus:

$$L_{\text{mes}} = \varepsilon L^0(T_{\text{obj}}) + (1-\varepsilon) L^0(T_{\text{env}}) \tag{4}$$

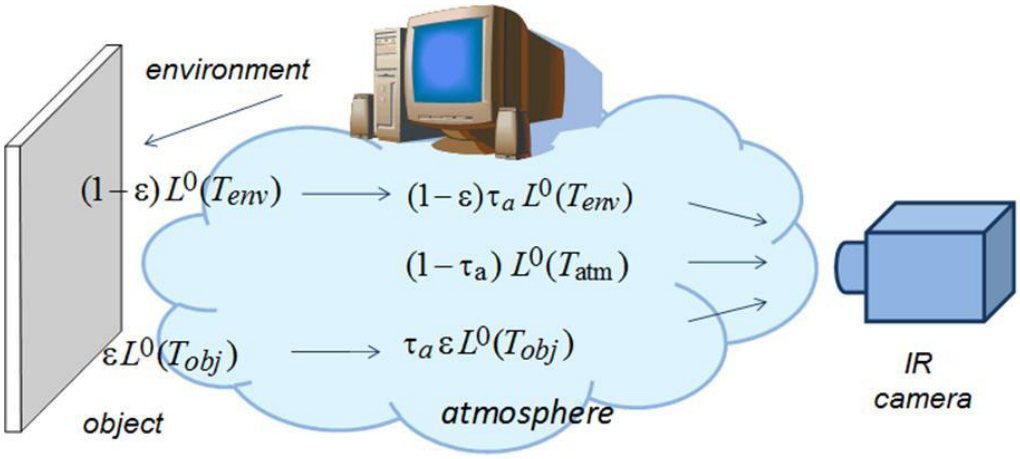


Figure 5 : *Simplified radiometric balance*

This equation shows that the environment must be reasonably well controlled in order to limit the influence of parasitic radiation (reflection from a radiator or any other radiative IR source, or even from the operator!). This precaution is all the more important when the measured temperature increases are minor. In addition, using a high emissivity coating (thus of low reflectivity) is obviously advantageous to minimize the parasitic flow/object flow ratio.

Along the same lines, note also the presence of the Narcissus effect (reflection of the cold detector on the scene), which is often observed when using a macro lens (e.g. lens magnification of 1, [8]). Usually, this is only an offset map which is superimposed on the scene, and which can thus be offset by subtraction of a reference image.

Last but not least, possible environmental instabilities could modify the exchange conditions between the sample and its environment and thus must be taken into account, especially when there are strong temperature variations over time.

4.3. Thermal calibration

In order to obtain reliable results, the user must, first of all, be confident in the apparatus calibration. Most of the time, infrared devices have their own setting and acquisition applications, including data-processing applications for digitizing, non-uniformity corrections, display, basic operations…

Generally, the calibration laws used by manufacturers suppose the sensor's response is linear, and consider the differences between the pixels' responses only as distributions of gains and offsets. The calibration of the device consists then in two distinct operations: the calibration of the average of a central area, and the application of maps of gains and offsets to link the response of each pixel to the one of the average of the sensor matrix. This second operation is called "Non-Uniformity Correction" (NUC).

The calibration law is generally taken in the form of a 2 or 3-degree polynomial, or a Planck-type law.

$$L_m(T) = aT^2 + bT + c \tag{5}$$

$$L_m(T) = \frac{R}{\left[\exp\left(\frac{B}{T}\right) - F\right]} + Offset \tag{6}$$

Where (*a*, *b*, *c*) or (*R*, *B*, *F* and *Offset*) are parameters identified during the calibration, and L_m the intensity measured by the camera, expressed in arbitrary units.

The gains and offsets maps are computed so as to obtain uniform distributions of digitized fluxes for two specific images of uniform thermal scenes taken at two different temperatures; these two scenes are generally obtained by means of an extended blackbody. Recently, some manufacturers proposed to go further, by linking the values of the gain and offset maps to an "internal temperature" of the camera, in order to compensate the thermal drifts associated with the heat produced by the internal electronics and the heat exchanges between the camera and its environment. This "advanced" non-uniformity correction is often called "Compensated Non-Uniformity Correction": CNUC.

Moreover, sensor matrixes always include some defective pixels (generally less than 0.5%), that can be saturated pixels, noisy pixels, or even "dead" pixels. They are localized using criteria dealing mainly with the discrepancy with respect to the mean response (in terms of digitized flux, gain, offset, etc.). Manufacturers propose to replace the value of these pixels by the one of their nearest non-defective neighbour (Bad Pixel Replacement, or BPR procedure), that induces a complete local correlation.

The validity of the standard calibrations can be easily checked out by observing a given thermal scene with a unique camera, but using different calibrations, associated with different acquisition settings (integration time / measurement range). As an illustration, figure 6 illustrates different observations on a blackbody using different infrared cameras. These two illustrations show that it is appropriate, if possible, to use the centre of the matrix and the middle of the calibration range when using the manufacturer's calibration laws. If the application needs a wider measurement area, it could be convenient to take into account the dispersion of the measured values in the data treatment procedure.

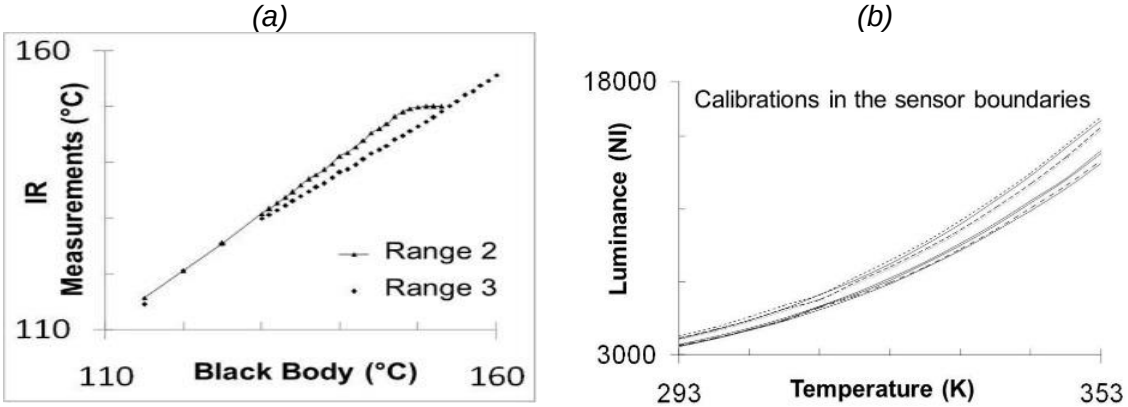


Figure 6 : *Check of the calibration using an extended black body: (a) Comparison between two ranges of a single camera (FLIR SC1000), (b) Comparison between several pixel responses (CEDIP IRC 320-4LW) of the array.*

If the specifications on the measurement accuracy are more stringent than one Kelvin, or if the independence of the measurement is a critical parameter for the later data processing, another solution is to be found. The most logical one consists in performing a customized calibration of the whole sensor matrix with testing conditions and camera configuration (integration time, windowing, etc.) similar to those used for the application, fitting the behaviour of each detector independently.

This calibration overcomes the limitations inherent to the NUC (or CNUC) and BPR procedures (linearity assumption valid further enough from saturation for the NUC, introduction of a strong spatial correlation between neighbouring pixels for the BPR operation…). However, it requires a high-uniformity extended blackbody so as to have a uniform radiation source at different temperature levels covering the whole range of the future application.

Once more, as in the standard global calibration procedure, the calibration law of each pixel can be chosen as a polynomial or as a Planck-like function, but the constant will be arrays of coefficients, the size of which being the one of the infrared matrix itself. These calibration coefficients are obtained by approximating, generally in the least squares sense, the couples (digitized radiation–temperature) by the chosen calibration function.

Defective pixels are then localized using a criterion for measuring the mismatch between the calibrated and specified temperature. The BPR operation is not performed: temperatures of the defective pixels are not taken into account in the subsequent data-processing. A specific pixel-to-pixel calibration is detailed in [9,10].

5. Conclusion

Accurate temperature measurement by radiative means is not an easy task. Many parameters have to be evaluated beforehand for extracting the surface emitted radiance from the measured radiance (atmospheric contributions: self-emission and attenuation, environments radiance reflections). One then faces the problem of temperature-emissivity separation. This underdetermined problem requires that some knowledge about the emissivity of the tested material is introduced. A general thought is that by adding spectral measurements at one or several other wavelengths would help identifying the temperature. The underdetermined nature of the problem is however maintained. Introducing a model for the emissivity spectral profile is often a misleading idea: high systematic errors unavoidably emerge when the model doesn't perfectly match to the real emissivity profile. Having some knowledge on emissivity *magnitude* helps much than imposing an arbitrary *shape* model.

References

[1] Gaussorgues G, "La Thermographie Infrarouge - Principes, Technologie, Applications", 4ème Édition, Tec & Doc Lavoisier, 1999 (in french)

[2] Pajani D, "Mesure Par Thermographie Infrarouge", Editeur ADD, EAN13 978293417107, 1989 (in french)

[3] Pron H, Menanteau W, Bissieux C, Beaudoin J-L, "Characterization of a focal plane array (FPA) infrared camera", QIRT 2000 (Eurotherm No. 64), Reims, 18-21 juillet 2000, pp 112-117

[4] Bissieux C, Pron H et Henry J-F "Pour de véritables caméras matricielles de recherche", revue Contrôles Essais Mesures, janvier 2003, vol n°2, pp. 39-41 (in french)

[5] Pron H, Laloue P, Henry J-F, L'ecolier J, Bissieux C et Nigon F, "Caractérisation de caméras infrarouges à matrice de détecteurs", 3ème Colloque Interdisciplinaire en Instrumentation, ENS Cachan, 29-30 janvier 2004, vol. 2, pp. 215-222 (in french)

[6] Pron H and Bissieux C, "Focal Plane Array infrared cameras as research tools", QIRT Journal, Vol. 1(2), 2004, pp. 229-240

[7] Datcu S, Ibos L, Candau Y, Mattéï S, Frichet J-C, "Focal plane array infrared camera transfer function calculation and image restoration", Optical Engineering, Vo. 43(3), 2004, pp. 648-657

[8] Poncelet M, Witz J-F, Pron H, Watrisse B, "A study of IRFPA camera measurement errors: radiometric artefacts", QIRT Journal. Vol. 8(1), 2011, pp. 3-20

[9] Honorat V, Moreau S, Muracciole J-M, B. Wattrisse B, Chrysochoos A, "Calorimetric analysis of polymer behaviour using a pixel calibration of an IRFPA camera", QIRT Journal, Vol. 2(2), 2005, pp.153-172

[10] Pron H, Bouache T, "Alternative thermal calibrations of Focal Plane Array infrared cameras", QIRT Journal, Vol. 13(1), 2016, pp 94-108

Lecture 5. Nonlinear parameter estimation problems: tools for enhancing metrological objectives

B. Rémy, S. André and D. Maillet[1]

[1] LEMTA, Université de Lorraine & CNRS, Vandœuvre-lès-Nancy, France

E-mails: benjamin.remy@univ-lorraine.fr
stephane.andre@univ-lorraine.fr
denis.maillet@univ-lorraine.fr

Abstract. The aim of this lecture is to present a methodology for enhancing the estimation of parameters in the case on a Non-Linear Parameter Estimation problem (NLPE). After some definitions and vocabulary precisions, useful tools to investigate NLPE problems will be introduced. Different techniques will be proposed for tracking for instance the true degree of freedom of a given estimation problem (Correlation, Rank of sensitivity matrix, SVD, ..) and enhancing the estimation of particular parameters by using either a Reduced model or a Model with some parameters fixed at their nominal values. The resulting reduced model can be unbiased or biased.

List of acronyms:

- **NLPE**: Non-Linear Parameter Estimation
- **PEP**: Parameter Estimation Problem
- **MBM**: Model-Based Metrology
- **SVD**: Singular Value Decomposition
- **OLS**: Ordinary Least Squares
- **SNR**: Signal-to-Noise Ratio

Scope

1. Introduction

The Non-Linear Parameter Estimation problem (NLPE) has been the subject of numerous lectures during the past METTI schools (see [1]). This text aims first at gathering in a synthetic way the basic notions and tools that can be used practically to analyse NLPE problems in engineering and science. At the same time, it provides new insights about the tools available to:

(i) enhance our knowledge about parameter identifiability in a given problem: which parameters can be really estimated in a given experiment and which precision can be achieved?
(ii) track the origin of pitfalls in parameter estimation problems (PEP),
(iii) offer new perspectives for enhancing the quality of model-based metrology (MBM) in a general way.

This lecture is composed of three different parts. The first one gives some definitions and vocabulary precisions. The second one presents some useful tools to investigate NLPE: ill-conditioned PEP will be considered and analysed and the use of SVD to track the PEP's degrees of freedom will be introduced next. The last part of this lecture consists in presenting some techniques for enhancing the performances of estimation, such as a dimensional analysis for identifying the degrees of freedom of a given problem and a reduction of the number of parameters involved in a theoretical model to make the PEP well-conditioned. As an example, the case of thermal characterization of a deposit on a substrate will be considered here.

2. Some definitions and vocabulary precisions

Performances of contemporary metrology, that is the science of measurement which includes material characterization for example, are not the result of the enhancement of the technology of measuring instruments only. They are also the consequence of the significant progresses accomplished in the field of Inverse Problems solving, especially when it is based on a very large amount of data. These are provided by new tools and by the facilities now available for numerical acquisition of experimental signals (CCD detectors allowing for 2D/3D numerical data acquisition and high frequency time resolution). Understanding the conditions for which parameters can be estimated from the model/measurements pair constitutes also a key point for reaching a high-quality estimation.

Measuring a physical quantity β_j requires a specific experiment allowing for this quantity to "express itself as much as possible" (notion of sensitivity). This experiment requires a system onto which inputs $u(t)$ are applied (stimuli) and whose outputs $y(t)$ are collected (observations). t is the explanatory variable: it corresponds to time for a purely dynamical experiment. A model M is required to mathematically express the dependence of the system's response with respect to quantity β_j and to other additional parameters β_k $(k \neq j)$: $\boldsymbol{y}_{mo} = \boldsymbol{\eta}(\boldsymbol{t}; \boldsymbol{\beta}, \boldsymbol{u})$ where input function $u(t)$ has been parameterized, that is decomposed under a finite set of basis functions, the coefficients of this decomposition being gathered in a vector $\boldsymbol{u}$ [8, page 26]. Many candidates may exist for function η - depending on the degree of complexity reached for modelling the physical process - which may exhibit different mathematical structure – depending for example on the type of method used to solve the model equations. Once this model is established, the physical quantities in vector $\boldsymbol{\beta}$ acquire the

status of model parameters. This model (called knowledge model if it is derived from physical laws and/or conservation principles) is initially established in a direct formulation. Knowing inputs $u(t)$ and the value taken by parameter $\boldsymbol{\beta}$, the output(s) can be predicted.

The linear or non-linear character of the model has to be determined:

- A Linear model with respect to its Inputs (LI structure) is such as:

$$y_{mo}(t; \boldsymbol{\beta}, \alpha_1 u_1(t) + \alpha_2 u_2(t)) = \alpha_1 y_{mo}(t; \boldsymbol{\beta}, u_1(t)) + \alpha_2 y_{mo}(t; \boldsymbol{\beta}, u_2(t)) \quad (1)$$

- A Linear model with respect to its parameters (LP structure) is such as:

$$y_{mo}(t; \alpha_1 \boldsymbol{\beta}_1 + \alpha_2 \boldsymbol{\beta}_2, u(t)) = \alpha_1 y_{mo}(t; \boldsymbol{\beta}_1, u(t)) + \alpha_2 y_{mo}(t; \boldsymbol{\beta}_2, u(t)) \quad (2)$$

In a metrological problem referred here as MBM (Model-Based Metrology), observations of the outputs will be provided by measurements. The inverse problem consists in making the direct problem work backwards with the objective of getting (extracting) $\boldsymbol{\beta}$ from $y_{mo}(t; \boldsymbol{\beta}, u(t))$ for given inputs and observations y. This is an estimation process. The difficulty stems here from two points:

(i) Measurements $\boldsymbol{y}$ are subjected to random perturbations (intrinsic noise $\boldsymbol{\varepsilon}$) which in turn will generate perturbed estimated values $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, even if the model is perfect: this constitutes an estimation problem.

(ii) the mathematical model may not correspond exactly to the reality of the experiment. Measuring the value of $\boldsymbol{\beta}$ in such a context leads to a biased estimation, where the bias is defined as $Bias = \mathrm{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}^{true}$, $\mathrm{E}(\hat{\boldsymbol{\beta}})$ being the expectation of the (stochastic) estimator $\hat{\boldsymbol{\beta}}$: this gives rise to an identification problem (which model structure η to use ?) associated to an estimation problem (how to estimate $\boldsymbol{\beta}$ for a given model structure?).

The estimation/identification process basically tends to make the model match the data (or the contrary). This is made by using some mathematical "machinery" aiming at reducing some gap (distance or norm)

$$\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{t}; \boldsymbol{\beta}, \boldsymbol{u}) \quad (3)$$

One of the obvious goals of NLPE (Non-Linear Parameter Estimation) studies is to assess the performed estimation through the calculation of the variances $V(\hat{\boldsymbol{\beta}})$ of the estimators of the different parameters. If the probabilistic distribution law of the noise is known, this allows to give the order of magnitude of confidence bounds for the estimates. NLPE problems require the use of Non-Linear statistics for studying such properties of the estimates.

Because of the two above-mentioned drawbacks of MBM, the estimated or measured value of a parameter β_j will be considered as "good" if it is not biased (or if its relative bias is low) and if its variance is minimum. Quantifying the bias and variance is also helpful to

determine which one of two rival experiments is the most appropriate for measuring the searched parameter (Optimal experiment design). In case of multiple parameters (vector $\boldsymbol{\beta}$) and NLPE problems, it is also interesting to determine which components of vector $\boldsymbol{\beta}$ are correctly estimated in a given experiment.

3. Useful tools to investigate NLPE problems

3.1. Sensitivities

The central role of the sensitivity matrix in PEP has been shown in the preceding lecture (Lecture 3). In the case of a single output signal $\boldsymbol{y}$ with m sampling points for the explanatory variable t and for a model involving n parameters, the sensitivity matrix is $(m \times n)$ defined as

$$S_{ij} = \left. \frac{\partial y_{mo}(t_i ; \boldsymbol{\beta}^{nom})}{\partial \beta_j} \right|_{t, \beta_k \text{ for } k \neq j} \tag{4}$$

As the problem is NL, the sensitivity matrix has only a local meaning. It is calculated for a given nominal parameter vector $\boldsymbol{\beta}^{nom}$.

If the model has a LP structure, this means that the sensitivity matrix is independent from $\boldsymbol{\beta}$. It can be expressed as (Lecture 3)

$$y_{mo}(t; \boldsymbol{\beta}) = \sum_{j=1}^{n} S_j(t)\, \beta_j \tag{5}$$

The sensitivity coefficient $S_j(t)$ to the j^{th} parameter β_j corresponds to the j^{th} column of matrix $\boldsymbol{S}$, once m discrete observation times have been chosen.

The primary way of getting information about the identifiability of the different parameters is to analyse and compare the sensitivity coefficients through graphical observations. This is possible only when considering reduced sensitivity coefficients $\boldsymbol{S}_j^*$ (sometimes called "scaled" sensitivity coefficients) because the parameters of a model do not have in general the same units.

$$\boldsymbol{S}_j^* = \beta_j \boldsymbol{S}_j = \beta_j \left. \frac{\partial \boldsymbol{y}_{mo}(\boldsymbol{t}; \boldsymbol{\beta}^{nom})}{\partial \beta_j} \right|_{\boldsymbol{t}, \beta_k \text{ for } k \neq j} = \left. \frac{\partial \boldsymbol{y}_{mo}(\boldsymbol{t}; \boldsymbol{\beta}^{nom})}{\partial (\ln \beta_j)} \right|_{\boldsymbol{t}, \beta_k \text{ for } k \neq j} \tag{6a}$$

Or

$$\boldsymbol{S}^* = \boldsymbol{S}\boldsymbol{R} \tag{6b}$$

with $\boldsymbol{R}$ the square diagonal matrix whose diagonal is composed of the components β_j of $\boldsymbol{\beta}$.

TOOL Nr1: A plot of all the reduced sensitivity coefficients $\boldsymbol{S}_j^*(\boldsymbol{t})$ gives a first idea about the most influential parameter for a given model (largest magnitude) and about possible correlations (sensitivity coefficients following the same evolution).

Example: Measurement of thermophysical properties of a coating layer through the Flash method using thermal contrast principle (Number of parameters $n = 2$).

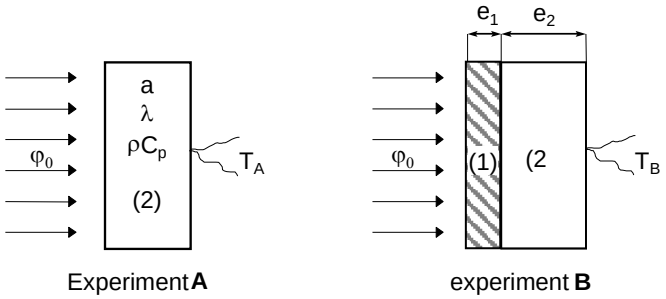


Figure 1 : *Basis of the "thermal contrast" method*

The thermal contrast method requires the repetition of two "flash" experiments A and B (**Figure 1**). The first one is operated on the substrate only (index (2)) whose thermophysical properties are known. The second experiment is performed on the two-layered sample (index (1)/(2)). In both cases, one records the rear face temperature evolutions. The thermograms so obtained are normalized with respect to their maximum and the difference of the scaled thermograms T_A and T_B is computed to produce the thermal contrast thermogram. This latter is a function of the thermophysical properties of the coating (1) and of the substrate (2) through two parameters:

$$K_1 = \frac{e_1}{e_2}\sqrt{\frac{a_2}{a_1}} \quad \text{and} \quad K_2 = \sqrt{\frac{\lambda_1 \rho_1 c_1}{\lambda_2 \rho_2 c_2}} \tag{7a}$$

The observable (contrast curve) and the reduced sensitivity coefficients to K_1 and K_2 are plotted in **Figure 2**. They show (i) that the sensitivities have the same order of magnitude as the signal (a good thing) but unfortunately (ii) these sensitivities appear to be totally correlated, since their maxima occur at roughly the same time (a bad thing). In this case, this simple plot shows that sensitivities to K_1 and K_2 are likely proportional and therefore that the identifiability of both parameters is impossible. This example will be more thoroughly modelled and studied in section 4 of this lecture.

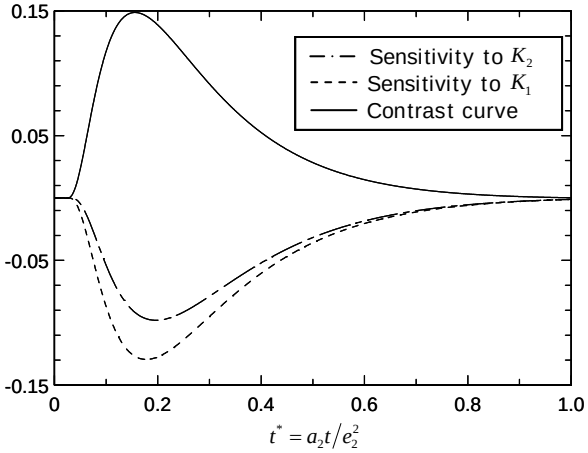


Figure 2 : *Reduced sensitivity coefficients for* $K_1 = 0.1$ *and* $K_2 = 1.36$

3.2. Variance/Correlation matrix

To go further and to investigate more deeply the PEP, the statistics of the estimator must be analysed. This can be made when (i) an estimator has been chosen (that is, a method to derive estimated values for the different parameters from the experimental signal), and (ii) the statistical properties of noise $\boldsymbol{\varepsilon}$ are known (according to experimentally founded observations).

We assume that the noise on the experimental signal is additive (this is in fact the definition of a noise), unbiased (which means that its stochastic average, its expectation is zero, for an unbiased model structure η of course) and independent (which means that the noise taken at two different times are independent) and has a constant variance σ^2 : this is sometimes called a IID. (Independent and Identically Distributed) noise, which occurs for perfect measurement with an ideal sensor. This corresponds to

$$y_i = y_{mo}(t_i ; \boldsymbol{\beta}) + \varepsilon_i \quad ; \qquad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \qquad ; \qquad \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_m \tag{7b}$$

where $\boldsymbol{I}_m$ is the identity matrix of size *m* (number of measurement points).

According to Beck's taxonomy (see [2] p. 134 and chapter VII), these assumptions correspond to the set "1111—11" with the following additional precisions: non stochastic independent explanatory variable (time), and no prior information for the parameters.

The OLS (Ordinary Least Squares) estimator $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes the least square sum, which gives:

$$J_{OLS}(\boldsymbol{\beta}) = \boldsymbol{r}^T(\boldsymbol{t};\boldsymbol{\beta},\boldsymbol{u})\,\boldsymbol{r}(\boldsymbol{t};\boldsymbol{\beta},\boldsymbol{u}) = \left\| \boldsymbol{r}(\boldsymbol{t};\boldsymbol{\beta},\boldsymbol{u}) \right\|^2 = \sum_{i=1}^{m} \left(y_i - y_{mo}(t_i;\, \boldsymbol{\beta},\boldsymbol{u}) \right)^2 \tag{8}$$

Where:

$$\boldsymbol{r}(\boldsymbol{t};\boldsymbol{\beta},\boldsymbol{u}) = \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{t};\boldsymbol{\beta},\boldsymbol{u}) \tag{9}$$

are defined as the residuals.

The estimator expression is found through a minimization process, where the j^{th} equation, also called "normal equation" is:

$$\partial J_{OLS}(t, \hat{\boldsymbol{\beta}}^{OLS}) \big/ \partial \beta_j = 0 \quad \text{for} \quad j = 1,2,...,n \tag{10a}$$

verified. If the global minimum of $J_{OLS}(\boldsymbol{\beta})$ is reached, the OLS estimator is unbiased, which means that the statistical mean of repeated estimated values $\hat{\boldsymbol{\beta}}$ is equal to the exact parameter vector $\boldsymbol{\beta}$.

Lecture 3 describes the behaviour of such an estimator for a LP model where the calculations can be fully completed to get an explicit linear OLS solution:

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\boldsymbol{S}^T \boldsymbol{S}\right)^{-1} \boldsymbol{S}^T \boldsymbol{y} \tag{10b}$$

In the case of a NL structure, the minimum is found through an iterative process using local linearity (Gauss-Newton algorithm basically, see [3]) of the form:

$$\hat{\boldsymbol{\beta}}_{OLS}^{(k+1)} = \hat{\boldsymbol{\beta}}_{OLS}^{(k)} + \left(\boldsymbol{S}^{(k)T} \boldsymbol{S}^{(k)}\right)^{-1} \boldsymbol{S}^{(k)T} \left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\hat{\boldsymbol{\beta}}_{OLS}^{(k)})\right) \tag{11}$$

The iterative process (12) requires computing the inverse of matrix $\boldsymbol{S}^T\boldsymbol{S}$ at each iteration *k*. Therefore, this latter must offer a good enough conditioning through repeated iterations. This is possible if the sensitivity coefficients are non-zero and linearly independent. Without any specialized and dedicated tool, this iterative process can be stopped when the residuals norm $\boldsymbol{r}^T\boldsymbol{r}$ is of the same order of magnitude as the measurement noise, that is when:

$$J_{OLS}(\hat{\boldsymbol{\beta}}^{(k)}) \approx m\ \sigma^2 \tag{12}$$

At convergence, the standard deviation of the error made for the estimated parameters can be evaluated thanks to the (symmetrical) **estimated** covariance matrix of the estimator. It characterizes the precision that can be reached on the estimated parameters (its inverse is sometimes named the precision matrix) and depends on the statistical assumptions that can be made on the data. In view of an OLS estimator, this matrix is

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) \equiv \begin{bmatrix} \operatorname{var}(\hat{\beta}_1) & \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2) & \cdots & \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_n) \\ \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2) & \operatorname{var}(\hat{\beta}_2) & & \operatorname{cov}(\hat{\beta}_2,\hat{\beta}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_n) & \operatorname{cov}(\hat{\beta}_2,\hat{\beta}_n) & \cdots & \operatorname{var}(\hat{\beta}_n) \end{bmatrix} \approx \sigma^2 \left(\boldsymbol{S}^T(\hat{\boldsymbol{\beta}})\, \boldsymbol{S}(\hat{\boldsymbol{\beta}})\right)^{-1} \tag{13}$$

It depends on the level of the Signal-to-Noise Ratio (SNR) and brings into play the inverse of the $\boldsymbol{S}^T\boldsymbol{S}$ matrix, already pointed out as a decisive operation for an accurate estimation. Matrix $\boldsymbol{S}^T\boldsymbol{S}$, which is also called the Fisher's information matrix with assumptions (8), depends on the number m of measurement points and on their distribution along the estimation interval, which may also be optimised if necessary [2]. The diagonal coefficients are the squares of the estimated standard deviation of each parameter $\sigma^2_{\hat{\beta}_j}$. They quantify the error that one can expect through inverse estimation. This is true if the assumptions made for the noise are

consistent with the experiment. The problem being NLP, retrieving these optimum bounds through a statistical analysis may depend on the starting guesses made to initialize the estimation algorithm. This matrix can also be an indicator for detecting possible correlations between the parameters. An estimation of the correlation matrix is calculated according to:

$$\mathbf{cor}(\hat{\boldsymbol{\beta}}) \approx \begin{bmatrix} 1 & \rho_{ij} & \cdots \\ \rho_{ij} & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \text{ all terms being the result of } \rho_{ij} = \frac{\mathrm{cov}(\hat{\beta}_i, \hat{\beta}_j)}{\sqrt{\sigma^2_{\hat{\beta}_i} \sigma^2_{\hat{\beta}_j}}} \tag{14}$$

The correlation coefficients (off-diagonal terms) correspond to a quantification of the 2 by 2 correlation existing between the two estimations of parameters β_i and β_j and, more precisely, between their errors (let us note that other forms of correlations involving more than 2 sensitivity coefficients exist, that is the multiple collinearity problem, which is detailed in section 3.3.2 further down). They vary between -1 and 1. They are global quantities (in some sense, "averaged" over the considered estimation interval, the whole $[0\ ,\ t_m]$ here). Gallant [4] suggested that difficulty in computation may be encountered when the common logarithm of the ratio of the largest to smallest eigenvalues of **cor** exceeds one-half the number of significant decimal digits used by the computer.

A more practical hybrid matrix representation **Vcor** can be constructed. It gathers the diagonal terms of the **cov**ariance matrix (more precisely their square root, normalized by the value of the estimated parameter) and the off-diagonal terms of the **cor**relation matrix.

$$\mathbf{Vcor}(\hat{\boldsymbol{\beta}}) \approx \begin{bmatrix} \sqrt{var(\hat{\beta}_i)} / \hat{\beta}_i & \rho_{ij} & \cdots \\ \rho_{ij} & \sqrt{var(\hat{\beta}_j)} / \hat{\beta}_j & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{15}$$

TOOL Nr2: Matrix Vcor $(\hat{\boldsymbol{\beta}})$ gives a quantitative point of view about the identifiability of the parameters. The main interest of this matrix lies in its diagonal coefficients, the relative standard deviation of the estimations of each parameter: these can be calculated independently from their physical units. These standard deviations of the estimated parameters are the stochastic root mean squares of the errors that are caused by the sole stochastic character of the IID noise, for an unbiased model.

The off-diagonal terms (correlation coefficients) are generally of poor interest because of their too global character. Values very close to ± 1 may explain very large variances (errors) on the parameters through a correlation effect**.**

NB: Another matrix, rcov $(\hat{\boldsymbol{\beta}})$ defined in equation (35) further on, is also very useful for assessing the quality of a potential inversion. Its diagonal coefficients are the squares of those of Vcor $(\hat{\boldsymbol{\beta}})$, but its off-diagonal coefficients are different.

Example: Here are two Vcor matrices taken from [1]. They were obtained for the same NLPE problems and for the same given set of nominal values of the $n = 3$ parameters but considering two different observables **A** and **B** (two different locations of the temperature measurements).

$$\mathrm{Vcor}_A(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.027 & 0.994 & -0.999 \\ \square & 0.0066 & -0.989 \\ \square & \square & 0.029 \end{bmatrix} \quad \text{Observable } \mathbf{A}$$

$$\mathrm{Vcor}_B(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.0002 & -0.38 & 0.63 \\ \square & 0.0008 & -0.93 \\ \square & \square & 0.0042 \end{bmatrix} \quad \text{Observable } \mathbf{B}$$

In the case of observable **A**, high relative standard deviations (nearly 3%) is observed for parameters β_1 and β_3 : it can be explained by a high degree of correlation between them ($|\rho_{13} = 0.999|$). Observable **A** can clearly not be used for estimating these parameters. On the contrary, observable **B** offers good identifiability for all parameters (small relative standard deviations) and does not show any 2 by 2 correlation.

3.3. Ill-conditioned PEP and strategies for tracking true degrees of freedom

3.3.1. Pathological example of ill-conditioning resulting from correlated parameters.

The good identifiability of parameters can be related to the local convexity of the cost functional $J_{OLS}(\boldsymbol{\beta})$ in the hyper-parameter space. One obvious consequence of a correlation between parameters is that several local minima may exist and make estimation algorithms consequently fail. The discussion that follows here is taken from an example of parameter estimation in a case of coupled radiative-conductive heat transfer [5]. The thermal characterization of a semi-transparent material implies a model depending on three basic parameters at least: the thermal diffusion characteristic time $t_d = e^2 / a$, the dimensionless optical thickness τ_0 and the dimensionless Planck number N (explanations to follow in section 4.1) and so $\boldsymbol{\beta} = [t_d, \tau_0, N]^T$. The estimation of the three parameters in this NLP problem may be difficult for some range of values of parameters τ_0 and N where matrix $\mathrm{Vcor}(\hat{\boldsymbol{\beta}})$ shows that a high degree of correlation between these two parameters exists, whereas the value of parameter t_d remains unconcerned.

A plot of the OLS criterium $J_{OLS}(\boldsymbol{\beta})$ in the 2D space (τ_0, N) for a given t_d value and a given noise σ (**Figure 3**) makes the consequence of such bad conditioning quite clear.

All level sets draw a very narrow valley oriented along a line which graphically corresponds to the relation $N \approx 2\,\tau_0$. A 3D plot would show that the central line of this valley does really correspond to a descending slope and hence that no real minima can be found. The level set indicated in the figure corresponds to exactly $J_{OLS}(\boldsymbol{\beta}) = 0.07 = m\,\sigma^2$. Trying to make the iterative optimization algorithm works below this limit for the stopping criterion is useless. In other words, the larger the noise, the higher the stopping level-set should be.

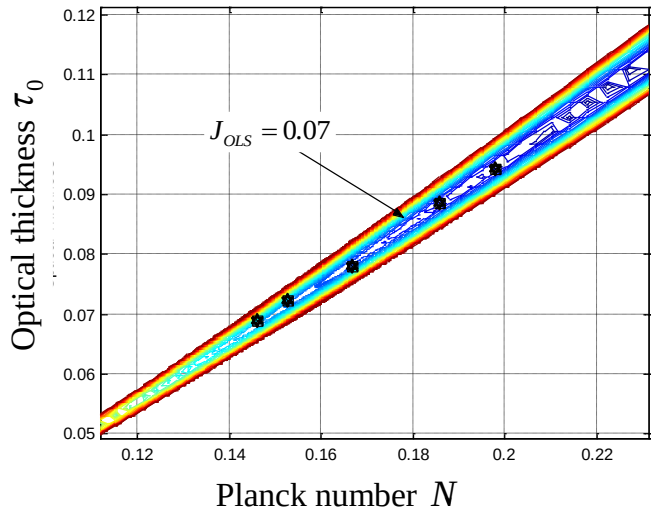


Figure 3 : *Level sets for $J_{OLS}(\boldsymbol{\beta})$ in the (τ_0, N) parameter space*

In the present case, this will not change the identifiability criterion. Depending on the initial guesses for the parameters, the deterministic algorithm will find different minima and different parameter estimates.

The four local minima are presented as big dots in **Figure 3** and correspond to the 3 parameters whose values are given in Table 1. Let us note that the local minimum Nr 4 Table 4) has been obtained with a stochastic algorithm (Simulated Annealing) different from a deterministic gradient based minimization algorithm used for finding the first 3 local minima. This shows that when the problem is ill-conditioned, stochastic algorithms are of little help for a correct estimation process (contrary to what is usually believed).

Such a behavior is more likely the result of a model which is not adapted to the physics involved. In the present case, it is interesting to note in Table 1 that all local minima that were found follows the relation $N(\tau_0 + 1)/\tau_0 = \text{Constant}$.

Parameter vector components	**Local Minima** *(found using either deterministic or stochastic algorithms)*			
	N°1	**N°2**	**N°3**	**N°4**
a *(10^7 m²/s)*	5.2	4.9	5.85	4.8
N	0.6	0.74	0.16	0.82
τ_0	0.38	0.5	0.076	0.56
$R_r = \frac{N_{Pl}}{\tau_0}(\tau_0 + 1)$	2.18	2.22	2.26	2.28

Table 1 : *Example of local minima found* $\hat{\boldsymbol{\beta}}$

In fact, an approximate modeling for conductive-radiative transfer in optically thin media can be shown to be more pertinent and more parsimonious. It makes naturally arise the notion of radiative resistance R_r which can be expressed as $R_r = N\,(\tau_0 + 1)/\tau_0$. This resistance is the appropriate parameter in this limiting behavior and prove that there is no way to identify independently τ_0 and N (Many different pairs are able to produce the same value for R_r).

TOOL Nr3: For an independent noise with known standard deviation and for a given model, it may be interesting to look at the level-set representation of the optimisation criterion in appropriate cut planes (for a given pair of parameters if *n* > 3) and compare it with the minimum achievable criterion given by $J = m\sigma^2$, where *m* is the number of measurements.

3.3.2. Rank of the sensitivity matrix.

We focus here on the scaled (or reduced) sensitivity matrix (see definition in equations (6a) and (6b)). This (*m*, *n*) matrix is composed of n column vectors, the reduced sensitivity coefficients $\boldsymbol{S}_j^*$:

$$\boldsymbol{S}^* = \left[\boldsymbol{S}_1^* \quad \boldsymbol{S}_2^* \quad \cdots \quad \boldsymbol{S}_n^*\right] \text{ with } \boldsymbol{S}_j^* = \beta_j \left.\frac{\partial \boldsymbol{\eta}\,(\boldsymbol{t};\boldsymbol{\beta}^{nom})}{\partial \beta_j}\right|_{\boldsymbol{t},\,\beta_k \text{ for } k \neq j} \tag{16}$$

where ***t*** is a column vector composed by all the *m* times of measurement:

$$\boldsymbol{t} = \left[t_1 \quad t_2 \quad \cdots \quad t_m\right]^T \tag{17}$$

These *n* column vectors $\boldsymbol{S}_j^*$ are in fact just the components of a set of *n* vectors $\vec{\boldsymbol{S}}_j^*$ in a *m*-dimension vector space. One can recall here that this set of vector $\Sigma = \{\vec{\boldsymbol{S}}_1^*, \vec{\boldsymbol{S}}_2^*, \ldots, \vec{\boldsymbol{S}}_n^*\}$ is linearly independent only if *m* coefficients α_j exist such as:

$$\sum_{j=1}^{n} \alpha_j \boldsymbol{S}_j^* = \boldsymbol{0} \Rightarrow \qquad \alpha_j = 0 \text{ for any } j \text{ with } 1 \leq j \leq n \tag{18}$$

This means that a linear combination of all these *m* vectors is equal to zero only if all its coefficients (the α_j's here) are equal to zero. If it is not the case, system Σ is linearly dependent. Let us note that the presence of a null vector in the set of vectors Σ makes it linearly dependent: such a null vector $\vec{\boldsymbol{S}}_j^*$ would correspond here to a parameter that has no influence on the variation of the model output, (the very specific case of a parameter β_j rigorously equal to zero is discarded here).

So, if the set is dependent, one has to remove one vector $\vec{\boldsymbol{S}}_j^*$ from the original set Σ and try again to test the independence condition (19) with the *n*-1 remaining vectors. This can be made

with the *n* possible choices for the vector $\vec{\boldsymbol{S}}_j^*$ that is removed from set Σ. If one finds one such independent set of *n*-1 vectors, the rank of the set is *n*-1. In the opposite case, one has to test the independence with *n*-2 vectors and so on... The rank *r* of Σ is the larger number of vectors for an independent subset of Σ that can be formed with the *n* original vectors.

In order to illustrate this, we will assume that $m = n = 2$ and that the model is linear. This corresponds to two observations of a model with two parameters β_1 and β_2. This leads to the set of two sensitivity vectors $\Sigma = \{\vec{\boldsymbol{S}}_1^*, \vec{\boldsymbol{S}}_2^*\}$ from which the situations shown in Figure 4 can be considered:

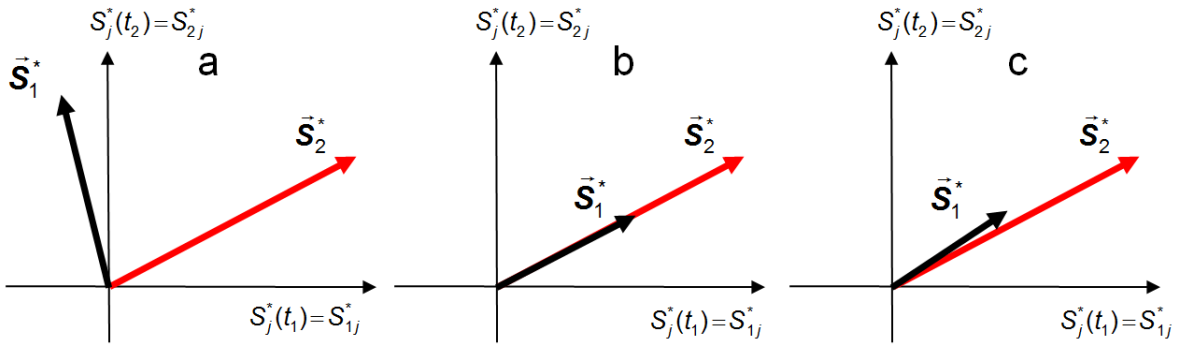


Figure 4 : *Reduced sensitivity vectors:*

a - independent sensitivities ($r = n = 2$) **b** - dependent sensitivities **c**- nearly dependent sensitivities

Case **a** corresponds to linearly independent sensitivity coefficients: the rank of Σ is equal to 2. It is also the rank of the reduced sensitivity matrix $\boldsymbol{S}^*$ and hence the rank of the sensitivity matrix, since $\boldsymbol{S}^* = \boldsymbol{S}\,\boldsymbol{R}$ (where $\boldsymbol{R}$ is the square diagonal matrix with two diagonal coefficients β_1 and β_2 according to equation 7). One can say that the observations of the model output provide two degrees of freedom since two parameters can be estimated.

Case **b** demonstrates a pathological nature of the sensitivity coefficients: they are proportional, with $\vec{\boldsymbol{S}}_2^* = 2\,\vec{\boldsymbol{S}}_1^*$ (one sees that the choice $\alpha_1 = 2$ and $\alpha_2 = -1$ in (19), which allows to show that the set of vectors Σ is not independent) and estimation of both coefficients is not possible anymore. In this case, the rank of $\boldsymbol{S}^*$ and hence the rank of $\boldsymbol{S}$ is $r = 1$ and the determinant of the information matrix $\boldsymbol{S}^T\boldsymbol{S}$ is equal to zero. This means that the explicit value of $\hat{\boldsymbol{\beta}}_{OLS}$, in the linear case (see equation 11b) and with a noise of spherical covariance matrix, which requires an inversion of the information matrix, is not possible. The same is true for the calculation of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$: the observations of the model output provide only one degree of freedom and only one parameter can be estimated, if the value of the other one is known.

Case **c** lies in between: the two reduced sensitivity vectors are nearly proportional $\vec{\boldsymbol{S}}_2^* \approx 2\,\vec{\boldsymbol{S}}_1^*$. Even if the mathematical rank is still equal to 2 (the previous equality is not an exact one), one

guesses that the number of degrees of freedom is somewhere between one and two and a more refined statistical analysis, taking into account the noise level in the measurements, has to be implemented.

Let us note that it is possible to test the presence of two nearly proportional vectors in set Σ, in the very general case, with of course a number of parameters less or equal to the number of observations ($n \leq m$), by testing the assumption $\vec{\boldsymbol{S}}_k^* - c_{kj}\vec{\boldsymbol{S}}_j^* = \text{or} \approx \vec{0}$, where c_{kj} is a proportionality constant: a plot of $S_k^*(t_i)$ as a function of $S_j^*(t_i)$, for the *m* common values t_i of the independent variable where observations are available (parametric representation of a curve) shows whether the plots gather on the $S_k^*(t) = c_{kj}\, S_k^*(t)$ line or not.

As an example of this type of representation, **Figure 5** illustrates the case taken from [1] of a 1D rear face transient response of a low insulating sample (conductivity λ) sandwiched between two very thin copper layers. The knowledge model (RDM1 in [1]) assumes pure thermal resistance for the insulating layer and pure known capacities for the copper layers. The front face is stimulated by a Dirac pulse of energy *Q* (J.m^{-2}), with a heat loss coefficient *h* (W.m^{-2} K^{-1}) equal over its two faces: the sensitivities to the three parameters *Q*, λ and *h* seem to be qualitatively independent, but only in terms of two by two linear dependencies: this does not mean that the rank of the reduced sensitivity matrix (if only these three parameters are looked for) is equal to three, because three by three linear dependencies may be possible.

This aspect, a possible dependency between the three sensitivity coefficients, is shown in **Figure 6**, for the same experimental design: a linear combination of the form $\vec{\boldsymbol{S}}_3^* - c_1\vec{\boldsymbol{S}}_1^* - c_2\vec{\boldsymbol{S}}_2^* = \text{or} \approx \vec{0}$ is looked for between the three sensitivity coefficients (for $\beta_1 = Q$, $\beta_2 = h$ and $\beta_3 = \lambda$) and a linear OLS estimation of c_1 and c_2 is made using the $S_1^*(t_i)$'s and the $S_2^*(t_i)$'s as the new independent variables and the $S_3^*(t_i)$'s as new observations. The corresponding $S_3^*(t_i)$ values are plotted as a function of the recalculated values (optimal linear combination) of the corresponding model, $\hat{c}_1\, S_1^*(t) + \hat{c}_2\, S_2^*(t)$: since the corresponding curve is very close to the first bisecting line, a *qualitative* 3 by 3 possible linear dependency is detected.

However, one can wonder how this dependency would impede the estimation of the three parameters: this has to be confirmed by a calculation of the covariance or Vcor matrix of the corresponding estimations, as explained in 3.2.

So, we will focus here on nonlinear parameter estimation problems where local linearization concepts as well as a Singular Value Decomposition of matrix deserve to be introduced.

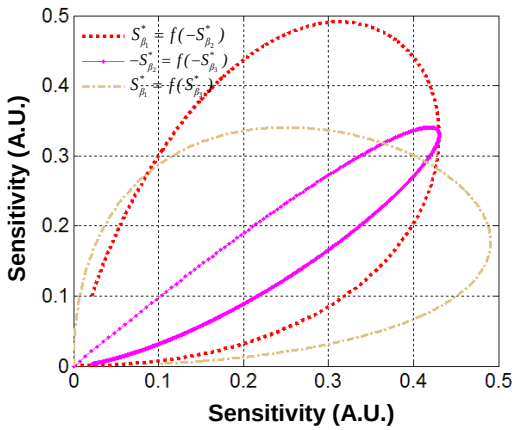


Figure 5 : *Sensitivities plotted by pairs*

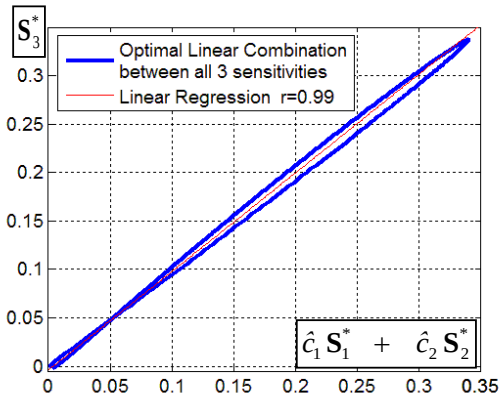


Figure 6 : *Evidence of Linear Combination between all three parameters*

3.3.3. Generalization: Use of SVD to track PEP degrees of freedom

It has been shown previously (see Lecture 3) that the question of identifiability of the parameters of a model relies on the condition number of the information matrix $\boldsymbol{S}^T\boldsymbol{S}$ if the physical units of the parameters are the same and of its scaled form $\boldsymbol{S}^{*T}\boldsymbol{S}^*$ if it is not the case. However, a systematic tool for tracking down hidden correlations is lacking. Such a tool will be presented now to circumvent this problem. Ultimately it will allow determining which parameters it is wise to exclude from the estimation (metrological) process, in order to get better estimates of the remaining ones.

In the next section two sequential steps will be presented.

First, in order to use all the tools available for linear estimation (see Lecture 3) on which the iterative OLS estimation (12) is based, the differential $\mathrm{d}\boldsymbol{y}_{mo}$ of the model will be calculated around a reference point $\boldsymbol{\beta}^{nom}$, that is a nominal value of the parameter vector for which a sensitivity analysis has been carried out (see previous sections) and the original parameter vector $\boldsymbol{\beta}$ will be made dimensionless using the components of $\boldsymbol{\beta}^{nom}$: a reduced parameter vector $\boldsymbol{x}$ with a well-defined norm will be constructed.

Second, Singular Value Decomposition (SVD) will be applied to the reduced sensitivity matrix of the "tangent" local linearized model around $\boldsymbol{\beta}^{nom}$, the ultimate goal being the determination the r parameters that can be estimated in a problem with n original parameters (with $n \geq r$), when the levels of the measurement noise and measurement magnitude are known (SNR).

The non-linear model $y_{mo}(\boldsymbol{t};\boldsymbol{\beta})$ is still considered here with m available measurements.

3.3.3.1. Parameterizing a non-linear parameter estimation problem around the nominal values of its parameters

The following single-output non-linear model is considered here:

$$y_{mo} = \eta\,(t;\boldsymbol{\beta}) \tag{19}$$

where $\boldsymbol{\beta}$ is the column vector of the n parameters, of size (n, 1), y_{mo} its (scalar) output at time t and η is a scalar function of t. If m observations of y_{mo} are available for times t_i, one can use a column vector notation:

$$\boldsymbol{y}_{mo} = \boldsymbol{\eta}\,(\boldsymbol{t};\boldsymbol{\beta}) \tag{20}$$

where $\boldsymbol{y}_{mo}$ is the output vector of the model, of dimensions (m, 1) and $\boldsymbol{t}$ the column vector of the m times of observation. In the relation, $\boldsymbol{\eta}\,(.)$ is a vector function whose values belong to R^m.

Since the model is non-linear, it will be written under a differential form, in the neighbourhood of a reference point $\boldsymbol{\beta}^{nom}$, which corresponds to a *nominal* value, where a sensitivity study has been already implemented. This allows to use a local linearity:

$$\mathrm{d}\boldsymbol{y}_{mo} = \boldsymbol{S}\,(\boldsymbol{t};\boldsymbol{\beta}^{nom})\,\mathrm{d}\boldsymbol{\beta} \qquad \text{with} \qquad S_{ij} = \left.\frac{\partial \eta\,(t_i;\boldsymbol{\beta}^{nom})}{\partial \beta_j}\right|_{\boldsymbol{t},\,\beta_k\ for\ k\neq j} \tag{21}$$

Let us note that in the notation $\mathrm{d}\boldsymbol{y}_{mo}$, the column vector $\boldsymbol{t}$ of the measurement times has been "frozen". $\boldsymbol{S}$ is the sensitivity matrix.

$$\boldsymbol{S} = \left[\boldsymbol{S}_1 \quad \boldsymbol{S}_2 \quad \cdots \quad \boldsymbol{S}_n\right] \quad \text{with} \quad \boldsymbol{S}_j = \left.\frac{\partial \boldsymbol{\eta}\,(\boldsymbol{t};\boldsymbol{\beta}^{nom})}{\partial \beta_j}\right|_{\boldsymbol{t},\,\beta_k\ \text{for}\ k\neq j} \tag{22}$$

In (22), the column vector $\mathrm{d}\boldsymbol{y}_{mo}$ has a norm, because all its m components have the same physical units. However, such is not the case for column vector $\mathrm{d}\boldsymbol{\beta}$, which is only a column matrix composed of n parameters whose physical dimensions are not necessarily the same: $\mathrm{d}\beta_1$ is a very small variation in the neighbourhood of β_1^{nom}, which can be a thermal conductivity λ. $\mathrm{d}\beta_2$ a very small variation around β_2^{nom}, which can be a volumetric heat capacity ρc and so on …

So $\mathrm{d}\boldsymbol{\beta}$ is not really a vector belonging to any vector space of dimension n, but a simple collection of n parameters.

In order to transform it into a real vector, a normalization of all its elements is necessary. The components of $\boldsymbol{\beta}^{nom}$ will be used for that purpose. A new dimensionless parameter $\boldsymbol{x}$ is introduced.

Its components are defined by:

$$x_j = \ln\left(\beta_j / \beta_j^{nom}\right) \tag{23}$$

And its nominal value is equal to zero:

$$\boldsymbol{x}^{nom} = \mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}^T \tag{24}$$

In the neighbourhood of $\boldsymbol{\beta}^{nom}$, each component of $\boldsymbol{x}$ is equal to the relative variation of the corresponding component of $\boldsymbol{\beta}$ around its nominal value (first order series expansion):

$$x_j = \ln\left(\beta_j / \beta_j^{nom}\right) = \ln\left(1 + \frac{\beta_j - \beta_j^{nom}}{\beta_j^{nom}}\right) \approx \frac{\beta_j - \beta_j^{nom}}{\beta_j^{nom}} \tag{25}$$

The new parameter vector $\boldsymbol{x}$ is written the following way:

$$\boldsymbol{x} = \ln\left(\boldsymbol{R}_{nom}^{-1}\,\boldsymbol{\beta}\right) \approx \boldsymbol{R}_{nom}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{nom}\right) \tag{26}$$

with :

$$\boldsymbol{R}_{nom} = \begin{bmatrix} \beta_1^{nom} & 0 & \cdots & 0 \\ 0 & \beta_2^{nom} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n^{nom} \end{bmatrix} \tag{27}$$

With this definition, the differential $\mathrm{d}\boldsymbol{x}$ of $\boldsymbol{x}$ is the logarithmic differential of $\boldsymbol{\beta}$:

$$\mathrm{d}\boldsymbol{x} = \begin{bmatrix} \mathrm{d}x_1 & \mathrm{d}x_2 & \cdots & \mathrm{d}x_n \end{bmatrix}^T \quad \text{with} \quad \mathrm{d}x_j = \frac{\mathrm{d}\beta_j}{\beta_j^{nom}} \approx \frac{\mathrm{d}\beta_j}{\beta_j} = \mathrm{d}\ln(\beta_j) \tag{28}$$

Let us note that the very last equality is only valid in the neighbourhood of $\boldsymbol{\beta}^{nom}$. It can also be written with a column vector notation:

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{R}_{nom}^{-1}\,\mathrm{d}\boldsymbol{\beta} \approx \boldsymbol{R}^{-1}\,\mathrm{d}\boldsymbol{\beta} \tag{29}$$

where $\boldsymbol{R}$ is the square diagonal matrix whose diagonal is composed of the components of $\boldsymbol{\beta}$, in the same way as (28) for the definition of $\boldsymbol{R}_{nom}$ starting from $\boldsymbol{\beta}^{nom}$.

Equation (22) is rewritten in order to make $\mathrm{d}\boldsymbol{x}$ appear:

$$\mathrm{d}\boldsymbol{y}_{mo} = \boldsymbol{S}^*\,\mathrm{d}\boldsymbol{x} \qquad \text{with} \quad \boldsymbol{S}^* = \boldsymbol{S}\,\boldsymbol{R}_{nom} \tag{30 a-b}$$

$\boldsymbol{S}^*$ is the reduced sensitivity matrix calculated for $\boldsymbol{\beta}^{nom}$, see (17, 23).

So, $\mathbf{d}\boldsymbol{y}_{mo}$ is a column vector belonging to R^m (it can be made truly dimensionless by a division by $\left\| \boldsymbol{\eta}\,(\boldsymbol{t};\boldsymbol{\beta}^{nom}) \right\|$ but it is not necessary here) and $\mathbf{d}\boldsymbol{x}$ is a true column vector belonging to R^m because its norm can be defined.

Using this change of variable as well as the SVD decomposition (*see Appendix 1*) of the scaled (also called reduced) sensitivity matrix $\boldsymbol{S}^*$, one can show that equation (31a) can be used to get a first order development in the neighbourhood of $\boldsymbol{\beta}^{nom}$ (see *Appendix 2* for the demonstration):

$$\boldsymbol{\beta} \approx \boldsymbol{R}_{nom}\left[\mathbf{1} + \boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T\left(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}\,(\boldsymbol{\beta}^{nom})\right)\right] = \boldsymbol{\beta}^{nom} + \boldsymbol{R}_{nom}\,\boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T\left(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}\,(\boldsymbol{\beta}^{nom})\right) \tag{31a}$$

with the following SVD decomposition: $\boldsymbol{S}^*\,(\boldsymbol{\beta}^{nom}) = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^T$ (32b)

Equation (12), that gives the Gauss-Newton algorithm can also be recast in terms of the scaled parameter $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{S}^{*T}\,(\boldsymbol{\beta}^{nom})\,\boldsymbol{S}^*\,(\boldsymbol{\beta}^{nom})\right)^{-1}\boldsymbol{S}^{*T}\,(\boldsymbol{\beta}^{nom})\left(\boldsymbol{y} - \boldsymbol{y}_{mo}\,(\boldsymbol{\beta}^{nom})\right) \tag{33}$$

This expression is equivalent to equation (12) where one has replaced the left-hand side $\boldsymbol{\beta}$ by its estimated value $\hat{\boldsymbol{\beta}}$ for a single iteration number k for $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)}$ and $\boldsymbol{\beta}^{nom} = \hat{\boldsymbol{\beta}}^{(k-1)}$. The complete demonstration is given in *Appendix 3.*

In a similar way, the variance-covariance matrix of scaled vector $\hat{\boldsymbol{x}}$ can be derived from (33) and (32b), see *Appendix 4*:

$$\mathrm{cov}(\hat{\boldsymbol{x}}) = \boldsymbol{R}_{nom}^{-1}\,\mathrm{cov}(\hat{\boldsymbol{\beta}})\left(\boldsymbol{R}_{nom}^{-1}\right)^T = \sigma^2\,\boldsymbol{V}\,\boldsymbol{W}^{-2}\boldsymbol{V}^T \tag{34}$$

One can note that, by definition, matrix $\mathrm{cov}(\hat{\boldsymbol{x}})$ is the reduced (or scaled) covariance matrix of $\hat{\boldsymbol{\beta}}$, which can be called rcov($\hat{\boldsymbol{\beta}}$):

$$\mathrm{rcov}(\hat{\boldsymbol{\beta}}) \equiv \mathrm{cov}(\hat{\boldsymbol{x}}) \equiv \begin{bmatrix} \sigma^2_{\hat{\beta}_1}/(\beta_1^{nom})^2 & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_2)/(\beta_1^{nom}\beta_2^{nom}) & \cdots & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_n)/(\beta_1^{nom}\beta_n^{nom}) \\ & \sigma^2_{\hat{\beta}_2}/(\beta_2^{nom})^2 & & \\ & & \ddots & \\ \text{Symmetric} & & & \sigma^2_{\hat{\beta}_n}/(\beta_n^{nom})^2 \end{bmatrix} = \sigma^2\left(\boldsymbol{S}^{*T}\boldsymbol{S}^*\right)^{-1} \tag{35}$$

One also shows, in Appendix 4, that the trace of $\mathrm{cov}(\hat{\boldsymbol{x}})$, that is the sum of the square of the relative standard deviations of all the estimations $\hat{\beta}_j$, at convergence, is equal to the sum of the square of the inverses of the singular values of $\boldsymbol{S}^*$, with a multiplicative factor equal to the variance σ^2 of the IID noise:

$$\mathrm{Tr}\left(\mathrm{cov}(\hat{\boldsymbol{x}})\right) \equiv \sum_{j=1}^{n}\left(\sigma_{\beta j} / \beta_j^{nom}\right)^2 = \sigma^2 \sum_{k=1}^{n} \frac{1}{w_k^2} \tag{36}$$

This allows to define a criterion m_q that assesses the global precision of the estimation:

$$m_q \equiv \left(\frac{1}{n}\sum_{j=1}^{n}\left(\sigma_{\beta j} / \beta_j^{nom}\right)^2\right)^{1/2} = \frac{1}{\sqrt{n}}\,\mathrm{Tr}\left(\mathrm{cov}(\hat{\boldsymbol{x}})\right) = \sigma\left(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{w_k^2}\right)^{1/2} \tag{37}$$

m_q is the root mean square relative standard deviation of the different parameters. So, it can be expressed in percent. If a specific parameter is estimated with a high relative variation, this will have an effect of m_q that will get large. The advantage of this criterion is that it takes into account the level of the measurement noise, contrary to the condition number of the relative sensitivity matrix $\mathrm{cond}(\boldsymbol{S}^*) \equiv w_1/w_n$ (see Lecture 3). It is quite easy to find an upper and a lower bound for it:

$$\frac{1}{\sqrt{n}}\frac{\sigma}{w_n} \le m_q = \left(\frac{1}{n}\sum_{j=1}^{n}\left(\sigma_{\beta j} / \beta_j^{nom}\right)^2\right)^{1/2} \le \frac{\sigma}{w_n} \tag{38}$$

Other points about this criterion that allows to study the well-posedness of a non-linear parameter estimation problem are given in *Appendix 4*.

TOOL Nr4: The SVD of the normalized sensitivity matrix calculated for nominal values of parameter vector $\boldsymbol{\beta}$ can bring valuable information to quantify the real identifiability of the parameters, once the level of noise known.

3.3.4 Residuals analysis and signature of the presence of a bias in the metrological process

One way to analyse the results of an estimation process is to calculate the residuals (equation 10) at convergence, when the assumptions (8) are fulfilled (an IID noise). When the model used for the estimation is not biased, the calculation of the residual column vector $\boldsymbol{r}(\hat{\boldsymbol{\beta}})$ whose k^{th} coefficients is the residual $r\ (t_k ; \hat{\boldsymbol{\beta}})$ at time t_k is:

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}) \equiv \boldsymbol{y} - \boldsymbol{y}_{mo}\ (\hat{\boldsymbol{\beta}}) = \boldsymbol{y}_{mo}\ (\boldsymbol{\beta}^{exact}) + \boldsymbol{\varepsilon} - \boldsymbol{y}_{mo}\ (\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\varepsilon} - \boldsymbol{S}\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{exact}) \text{ with } \boldsymbol{S} = \boldsymbol{S}\ (\hat{\boldsymbol{\beta}}) \tag{39}$$

One shows in *Appendix 5* that, strictly speaking, the residuals, when the model is unbiased, are correlated but, in practice, adding more measurements times for a given estimation interval tends to make them nearly uncorrelated. This is especially true for thermal characterization of materials or systems, where the number of parameters is low (2, 3, 4, ...) and the time sampling

rate high enough with respect of the length of measurement (several hundredth of measurements at least for modern data acquisition systems).

So, when these previous conditions are fulfilled, "signed" residuals can be considered as the signature of some estimation based on a biased model.

This bias can stem from different causes such as:

(i) the a priori decision that some parameters of the model are known and therefore fixed at some given value (maybe measured by another experiment). As active parameters in the PEP, they can alter the estimates of the remaining unknown parameters.

(ii) Experimental imperfections which make the model idealized with respect to the reality of the phenomena.

The existence of a bias means that a systematic and generally unknown inconsistency exists between the model and the experimental data.

We give here an example taken from [1] and already studied in section 3.3.2 above. It concerns the simulation of a flash experiment applied to a three-layer medium: two highly capacitive and conductive coatings and a central layer made of a material with very poor conductivity (highly insulating material) and heat capacity (aerogel material). This system can be modelled through some function $T^{rear} = y_{mo}(t, \boldsymbol{\beta})$. An artificial bias $d(t)$ is introduced under the form of a linear drift superimposed to the output simulated observations. It corresponds practically to a linear deviation of the signal from the equilibrium situation before the experiment starts. So, the correct model that should be used to mimic the observed rear face measurement should be:

$$y_{mo}^{drift}(t_k, \boldsymbol{\beta}^{exact}) = y_{mo}(t_k; \boldsymbol{\beta}^{exact}) + d(t_k) \tag{40}$$

A noise respecting equation (8) is also added to the simulation of the measurements so that we have at each time t_k:

$$y_k = y_{mo}^{drift}(t_k, \boldsymbol{\beta}^{exact}) + \varepsilon_k \tag{41}$$

Of course, model $y_{mo}(t, \boldsymbol{\beta})$ is exact if no drift is present in the experiment. However, in the opposite case, it becomes biased, since it does not account for the presence of this drift.

Let us note that in this definition, the drift model is the reference one ($y_{mo}^{exact} = y_{mo}^{drift}$) and the preceding thermal model is the biased one ($\boldsymbol{y}_{mo}^{biased} = \boldsymbol{y}_{mo}$).

If this biased model is used for estimation, the minimization will be done by a minimization of the following criterion based on a biased residual vector:

$$J_{biased}(\boldsymbol{\beta}) = \boldsymbol{r}_{biased}^T(\boldsymbol{\beta})\, \boldsymbol{r}_{biased}(\boldsymbol{\beta}) \quad \text{with} \quad \boldsymbol{r}_{biased}(\boldsymbol{\beta}) \equiv \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{y}_{mo}^{drift}(\boldsymbol{\beta}) + \boldsymbol{d} \tag{42}$$

As a consequence, at convergence, the error on the estimated parameters vector will have a deterministic part and a stochastic part:

$$\boldsymbol{e}_\beta \equiv \hat{\boldsymbol{\beta}}^{biased} - \boldsymbol{\beta}^{exact} \approx \boldsymbol{b}_\beta + \boldsymbol{A}\boldsymbol{\varepsilon} \quad \text{with } \boldsymbol{b}_\beta = \boldsymbol{0} \text{ if } \boldsymbol{d} = \boldsymbol{0} \tag{43}$$

where $\boldsymbol{A}$ is a matrix that corresponds to the linearization of the inverse problem with respect to the noise in the neighbourhood of the exact value of $\boldsymbol{\beta}^{exact}$ and $\boldsymbol{b}_\beta$ a bias of non-zero average, that stems from the presence of the drift $\boldsymbol{d}$.

As a consequence, the residual defined in (42) can be calculated, at convergence, using (43):

$$\boldsymbol{r}_{biased}(\hat{\boldsymbol{\beta}}^{biased}) \equiv \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}) = \boldsymbol{y}_{mo}^{drift}(\boldsymbol{\beta}^{exact}) + \boldsymbol{\varepsilon} - \boldsymbol{y}_{mo}(\hat{\boldsymbol{\beta}}^{biased}) \tag{44}$$

or

$$\boldsymbol{r}_{biased} = \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact}) + \boldsymbol{d} + \boldsymbol{\varepsilon} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact} + \boldsymbol{b}_\beta + \boldsymbol{A}\boldsymbol{\varepsilon}) \tag{45}$$

A first order development of the last term around the exact value $\boldsymbol{\beta}^{exact}$ yields:

$$\boldsymbol{r}_{biased}(\hat{\boldsymbol{\beta}}^{biased}) = \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact}) + \boldsymbol{d} + \boldsymbol{\varepsilon} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact}) - \boldsymbol{S}(\boldsymbol{\beta}^{exact})\left[\boldsymbol{b}_\beta + \boldsymbol{A}\boldsymbol{\varepsilon}\right] \tag{46}$$

or

$$\boldsymbol{r}_{biased}(\hat{\boldsymbol{\beta}}^{biased}) = \boldsymbol{d} + \boldsymbol{S}(\boldsymbol{\beta}^{exact})\ \boldsymbol{b}_\beta + \left[\boldsymbol{I} - \boldsymbol{S}(\boldsymbol{\beta}^{exact})\boldsymbol{A}\right]\boldsymbol{\varepsilon} \tag{47}$$

This means that the residuals are biased, because of their first deterministic component, even if its second stochastic one may be diagonal.

We return here to the estimation problem described in section 3.3.2 (flash experiment on a three layer sample for the inner insulating layer characterization): we have seen that the model used for parameter estimation was ill-conditioned: some correlation exists between the parameters (Case $n = 3$ corresponding to the correlation existing between parameters shown in **Figure 5** and **Figure 6**). **Figure 7** below shows that

- the simulated rear face noisy output of the system, with the drift and some added noise (dotted curve),
- the corresponding rear face recalculated output using the biased estimate $\hat{\boldsymbol{\beta}}$ (obtained through minimization of criterion (42)) - (blue solid line),
- the drift of the model output (function $b_y(t)$) introduced (brown solid line) . At the final time of the experiment ($t_f = 1000$ s), the magnitude of the drift represents less than 4% of the maximum level of the signal,
- the residuals curve, with the noised signal (minimization of criterion (42), grey stochastic line), and after removing the noise, that is with the same estimation process starting from a noiseless signal, that is with $\boldsymbol{\varepsilon} = \boldsymbol{0}$, blue solid line).

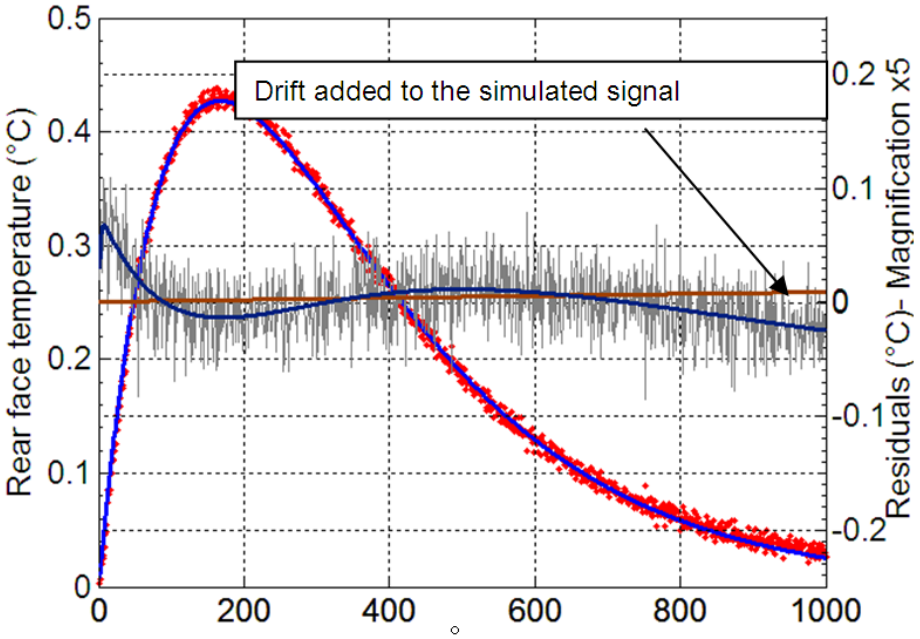


Figure 7 : *Signed character of "post-estimation" residuals in the presence of a bias and using a badly conditioned PEP*

The "signed" character of the residuals is obvious (oscillation around zero with a much smaller frequency than the noise). The three parameters estimated (Q, h and λ) using these biased "measurements" have averaged values (obtained by repeated Monte Carlo simulated measurements) that differ respectively by -18%, -7.5%, +19% from the exact input values. These differences are not of stochastic origin (caused by noise only) but result from the introduction of the bias. One possibility for the experimenter who wants to check whether his estimations are biased or not, is to observe the output of the inversion process for varying identification ranges of the independent variable. For example, we can vary the identification time interval. If a bias affects the data when compared to the modeling, then the estimations will vary, depending on the selected identification interval. This is what can be observed in **Table 2** where three identifications have been performed for three different time intervals [0-70s], [0-150s], [0-300s]. In this case we have used a more refined model than the one used for **Figure 7** and thus a more badly conditioned PEP. In this table both thermal properties of the insulating material (thermal conductivity and thermal diffusivity) were estimated from the biased data. Obviously with such a material, the small heat capacity makes a good estimation of this parameter difficult, but sadly (because of a lack of sensitivity) this also affects the estimation of the second parameter. The thermal diffusivity and conductivity estimated from the data of **Figure 7** depend strongly on the identification intervals. The values can change within a factor of 60% or 170% in that case.

Time Interval	70 s	150 s	300 s
a (m²/s)	$3.76.10^{-6}$	$3.22.10^{-6}$	$2.21.10^{-6}$
λ (W/m/°C)	0.031	0.064	0.084

Table 2 : Influence of the existence of some bias on the parameter estimates for a badly conditioned problem

TOOL Nr5: The "post-estimation" residuals have to be analysed carefully to check the potential existence of a bias of systematic origin. Its magnitude can be compared to the standard deviation of the white noise of the sensor in order to check whether this bias may introduce too large confidence intervals for the estimates (with respect to the pure stochastic estimation of the variances of parameter estimates in the absence of any bias). Invariant estimates for different identification intervals suggest that the bias is acceptable. In the opposite case, strategies must be implemented, either to change the nature of the estimation problems (reduction of the initial goals) or to use residuals to give a fair quantitative evaluation of the confidence bounds of the estimates. Some hints on that topic will be given in the next section.

4. Enhancing the performances of estimation

Some tools have been given above: they can help the experimenter to gain insight into its metrological problem. They can lead to a conclusion of failure: the problem is ill-conditioned regarding the estimation of the interesting parameters. This means that the parameters we initially wish to measure will actually never be estimated accurately. Two strategies are possible: recognizing that the initial goal is in vain or modifying the problem through physical thinking to make it well-posed or adequately conditioned even by changing the goals themselves (number of parameters to estimate). Quoting J.V.Beck [2]: "the problem of non-identifiability can be avoided, through either the use of a different experiment or a smaller set of parameters that are identifiable".

This position emerges from the well-known parsimony “principle” (see http://en.wikipedia.org/wiki/Parsimony) which in the field of science could be summarized by this sentence : “trying to perfectly recover reality is indeed very easy, when one adds parameters to each other so that it connects-the-dots”. There is much more to learn and to retrieve from the distance maintained between a model and the observations it is supposed to match. The resulting consequence is that any minimization algorithm is a good one because the problem is well defined. This section will now proceed to give additional tools to work out badly conditioned problems with special analysis regarding the role of known versus unknown parameters.

4.1 Dimensional analysis or natural parameters: case of coupled conduction/radiation flash experiment

Through the preceding sections, the reader should have been convinced of the importance of notions like the pertinence of a model (good representation of reality, controlled origins of bias), the application of the parsimony principle that is to adapt one's metrological objective by making the "quality" of the available information match the degree of complexity of the model.

A reduced model, seen as a model with a reduced number of parameters, has to be considered first in the light of Dimensional Analysis. The principles of Dimensional Analysis in Engineering precisely rely on the construction of "appropriate" natural parameters (the Pi-groups) emerging from the rank determination of the dimensional matrix of all physical quantities involved in the problem with respect to a basis of "base" quantities [6].

If we consider the heat transfer problem in a semi-transparent material like glass, coupled conduction and radiation transfers must be considered. Material parameters involve classical

thermophysical properties of the opaque material (thermal conductivity λ, specific heat ρc) with the additional parameters accounting for radiative transfer : the absorption (extinction coefficient) β (m^{-1}), the level of temperature of the material T_0 (in Kelvin) which rules the magnitude of radiation emission, the Stefan-Boltzmann constant σ_{SB}, the refractive index n, and the inner emissivity ε_i of the boundaries (no units - opaque coatings of the glass slab are considered here).

Let us assume that a flash experiment is planned, with an absorbed heat density Q ($J.m^{-2}$). In order to study the possibilities for a transient thermal characterization technique of such materials (which parameters can be measured with this experiment?), the model will give the rear face temperature response of the slab (thickness e) as the following function:

$$y_{mo} = T_{rear}^{flash}\ (t, e, Q, \rho c, \lambda, \beta, \sigma_{SB}, T_0, \varepsilon_i, n) \tag{48}$$

Practicing a "blind" Dimensional Analysis leads to the construction of a new function depending on a new set of parameters:

$$y_{mo} \equiv \frac{T_{rear}^{flash} - T_0}{T_0^*} = T_{rear}^{flash*}\left(t^* = \frac{at}{e^2}, \tau_0 = \beta e, N = \frac{\lambda \beta}{n^2 \sigma_{SB} T_0^3}, T_0^* = \frac{Q}{\rho c e}, \varepsilon_i\right) \tag{49}$$

which naturally produces 4 pi-groups governing heat transfer inside the sample, with a reduction of the number of initial parameters of the model from 10 to 5.

Another classical example deals with conductive and convective mechanisms of transfer which appear jointly in problems of heat transfer within boundary layers. Solving the Inverse Heat Conduction Problem in order to get a heat exchange coefficient estimation will require the introduction of the classical Reynolds, Nusselt and Prandtl numbers.

4.2 Reducing the PEP to make it well-conditioned: case of thermal characterization of a deposit

- **Model:** *Case of the contrast method*

The method of the thermal contrast already presented in **Section 3.1** consists in making two "flash" experiments in order to estimate the thermal properties of the coating layer, denoted (1) in **Figure 8** below (the same as **Figure 1**). We will now on detail the modelling already presented briefly in section 3.1, in order to be able to find out which parameters of the model can be really estimated, in this non-linear parameter estimation problem.

Let us remind that the first flash experiment is carried out on the substrate denoted (2), which allows characterization of the substrate in terms of diffusivity (the thermal capacity of the substrate is measured by another facility). The second flash experiment is performed on the two-layer material denoted (1)/(2).

In both cases, the variation of the rear-face temperature *T* with time, called thermogram, is measured. By taking the difference of theses thermograms T_A^* and T_B^* normalized by their respective maximum, we obtain a curve called a thermal contrast curve, which is a function of the thermophysical parameters of the film (1) and of the substrate (2).

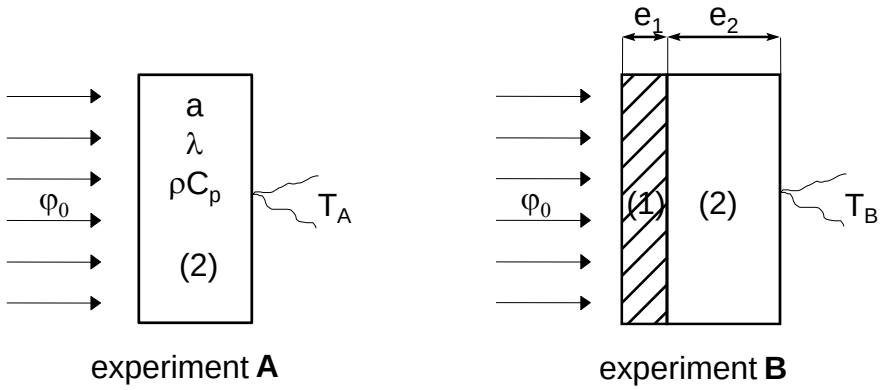


Figure 8 : *Principle of the Method*

The thermal quadrupoles method [7] is very appropriate to find the rear-face temperatures. Taking the Laplace transform of the heat equation yields a linear relationship between the different quantities of the "in" and "out" faces of each layer of the material.

Let $\theta(z,p)$ and $\phi(z,p)$ being the Laplace transforms of the temperature $T(z,t)$ and heat density $\varphi(z,t)$ respectively, with z the axis normal to both faces:

$$\theta(z,p) = \mathsf{L}\left[T(z,t)\right] = \int_0^\infty T(z,t)\exp(-pt)\mathrm{d}t \tag{50}$$

and

$$\phi(z,p) = \mathsf{L}\left[\varphi(z,t)\right] = \int_0^\infty \varphi(z,t)\exp(-pt)\mathrm{d}t \qquad \text{with} \qquad \varphi(z,t) = -\lambda\frac{\partial T}{\partial z} \tag{51}$$

The thermal quadrupoles method allows to linearly link the temperatures and the heat flux densities of a homogeneous layer (numbered *i* here) without any source term and with zero initial temperature, through a transfer matrix M_i, defined in the following way:

$$\begin{bmatrix}\theta_{iin}\\ \phi_{iin}\end{bmatrix} = \begin{bmatrix}A_i & B_i\\ C_i & D_i\end{bmatrix}\begin{bmatrix}\theta_{iout}\\ \phi_{iout}\end{bmatrix} \tag{52}$$

with the coefficients of the matrix being calculated as:

$$A_i = D_i = \cosh\left(e_i\sqrt{p/a_i}\right);\ B_i = \frac{1}{\lambda_i\sqrt{p/a_i}}\sinh\left(e_i\sqrt{p/a_i}\right);\ C_i = \lambda_i\sqrt{p/a_i}\ \sinh\left(e_i\sqrt{p/a_i}\right)$$

The subscript (i) is related to the layer (i) : film (1) and substrate (2).

e_i : thickness of the material
a_i : thermal diffusivity
λ_i : thermal conductivity
ρC_{p_i} : specific heat

It is convenient in this 1D transient problem, to notice that time can be made dimensionless with the thermal diffusivity a_2 of the substrate and with its thickness e_2, to make a Fourier number t^* appear, which will be associated to a reduced Laplace parameter p^* defined as:

$$t^* = \frac{a_2 t}{e_2^{\,2}} \ , \ p^* = p \frac{e_2^{\,2}}{a_2} \text{ and } \ s = \sqrt{p^*} \tag{53}$$

We can then define a reduced Laplace transform $\tilde{\theta}$ as:

$$\tilde{\theta}(z,p*) = \tilde{\mathrm{L}}\left[T\left(z,t^*\right)\right] = \int_0^\infty T(z,t*)\exp(-p*t*)\mathrm{d}t* = \frac{a_2}{e_2^2}\,\theta(z,p) \tag{54}$$

- **Flash Experiment on the substrate:**

The expression of the rear face response to a pulsed (Dirac) stimulation $\varphi(t) = Q_2\,\delta(t)$, where Q_2 is the energy density (in J.m^{-2}) absorbed by the front face, is given by the following relationship:

$$\begin{bmatrix} \theta_{2in} \\ \phi_{2in} = Q_2 \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} \theta_{2out} \\ \phi_{2out} = 0 \end{bmatrix} \tag{55}$$

Hence:
$$\theta_{2out} = \frac{Q_2}{C_2} = \frac{Q_2}{\lambda_2 \sqrt{\frac{p}{a_2}} \sinh\left(\sqrt{\frac{p e_2^{\,2}}{a_2}}\right)} \tag{56}$$

Here subscript 'in' designates the front (stimulated) face while subscript 'out' is associated to the rear face, where temperature can be measured. This rear face is supposed to be insulated here ($\phi_{2out} = 0$ in (55)).

Setting $s = \sqrt{p^*}$ and normalizing the thermogram with respect to its maximum that corresponds to the adiabatic temperature: $T_{2_\infty} = \frac{Q_2}{\rho_2 c_2 e_2}$ reached for long times for this adiabatic model, we obtain:

$$\theta^*_{2_{out}} = \mathsf{L}\left(\frac{T_2}{T_{2_\infty}}\right) = \frac{e_2^{\,2}}{a_2}\frac{1}{s\sinh(s)} \tag{57}$$

Using the reduced Laplace transform (57), we can write:

$$\tilde{\theta}^*_{2_{out}} = \tilde{\mathsf{L}}\left(\frac{T_2}{T_{2_\infty}}\right) = \frac{1}{s\sinh(s)} \tag{58}$$

- **Flash Experiment on the two-layer material:**

The expression of the rear face response of the two-layer material can also be obtained easily through the quadrupoles method:

$$\begin{bmatrix} \theta_{1/2in} \\ \phi_{1/2in} = Q_{1/2} \end{bmatrix} = \begin{bmatrix} A_{eq} & B_{eq} \\ C_{eq} & D_{eq} \end{bmatrix} \begin{bmatrix} \theta_{1/2out} \\ \phi_{1/2out} = 0 \end{bmatrix} \tag{59}$$

where:
$$\begin{bmatrix} A_{eq} & B_{eq} \\ C_{eq} & D_{eq} \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} = \begin{bmatrix} A_1A_2 + B_1C_2 & A_1B_2 + A_2B_1 \\ A_1C_2 + A_2C_1 & A_1A_2 + B_2C_1 \end{bmatrix} \tag{60}$$

and where $Q_{1/2}$ is the energy density absorbed by the front face in this second flash experiment on the two-layer sample.

In the case of good conductive materials with small thicknesses, the Biot number which represents the ratio between the internal resistance and the external resistance is low, which justifies neglecting the heat losses in the model output (rear face temperature) above. The expression of the temperature takes the following form:

$$\theta_{1/2} = \frac{Q_{1/2}}{C_{eq}} = \frac{Q_{1/2}}{A_1C_2 + A_2C_1} \tag{61}$$

Note: If we switch the two layers of the material, it means inverting subscripts 1 and 2, and the expression of the rear-face temperature can be proved to remain unchanged.

$$\theta_{1/2out} = \frac{Q_{1/2}}{\lambda_1\sqrt{\frac{p}{a_1}}\sinh\left(\sqrt{\frac{pe_1^{\,2}}{a_1}}\right)\cosh\left(\sqrt{\frac{pe_2^{\,2}}{a_2}}\right) + \lambda_2\sqrt{\frac{p}{a_2}}\sinh\left(\sqrt{\frac{pe_2^{\,2}}{a_2}}\right)\cosh\left(\sqrt{\frac{pe_1^{\,2}}{a_1}}\right)} \tag{62}$$

If we now scale the thermogram with the adiabatic temperature of the two-layer material, that is with $T_{1/2_\infty} = \dfrac{Q_{1/2}}{\rho_1 c_1 e_1 + \rho_2 c_2 e_2}$, the expression of the Laplace transforms of this reduced temperature temperature $T_{1/2} / T_{1/2\infty}$ takes a simpler form:

$$\theta^*_{1/2_{out}} = \frac{e_2^2}{a_2} \frac{1 + \dfrac{\rho_1 c_1 e_1}{\rho_2 c_2 e_2}}{s\left[\sqrt{\dfrac{\lambda_1 \rho_1 c_1}{\lambda_2 \rho_2 c_2}}\, sinh\left(\dfrac{e_1}{e_2}\sqrt{\dfrac{a_2}{a_1}}s\right) cosh(s) + sinh(s) cosh\left(\dfrac{e_1}{e_2}\sqrt{\dfrac{a_2}{a_1}}s\right)\right]} \tag{63}$$

As in section 3.1 two reduced parameters are introduced:

$K_1 = \dfrac{e_1}{e_2}\sqrt{\dfrac{a_2}{a_1}}$: ratio of the root of characteristic times

or $K_1 = \sqrt{tc_1 / tc_2}$ with $tc_i = e_i^2 / a_i$ for $i = 1, 2$ (64)

$K_2 = \sqrt{\dfrac{\lambda_1 \rho_1 c_1}{\lambda_2 \rho_2 c_2}}$: ratio of the thermal effusivities

or $K_2 = \sqrt{b_1 / b_2}$ with $b_i = \sqrt{\lambda_i \rho c_i}$ for $i = 1, 2$ (65)

We can note that K_1 is a function of the thicknesses of the substrate and coating and K_2 is an intrinsic parameter of the materials. The reduced Laplace transform of the response of the two-layer system can then be written, using (54):

$$\boxed{\tilde{\theta}^*_{1/2_{out}} = \frac{1}{s}\left[\frac{1 + K_1 K_2}{K_2 \sinh(K_1 s)\cosh(s) + \sinh(s)\cosh(K_1 s)}\right]} \tag{66}$$

The heterogeneous nature of the two-layer material system appears here through the expression of the denominator that cannot be simplified: this makes the definition of an equivalent material associated to this two-layer sample impossible.

- **Contrast Curve:**

The contrast curve is obtained by taking the difference between the two thermograms, that is:

$$\Delta\tilde{\theta}^*_{out} = \tilde{\theta}^*_{1/2_{out}} - \tilde{\theta}^*_{2_{out}} = \tilde{\mathcal{L}}\left(T^*_{1/2_{out}} - T^*_{2_{out}}\right) = \tilde{\mathcal{L}}\left(\Delta T^*\right) \tag{67}$$

The expression of the reduced thermal contrast in the Laplace domain is:

$$\boxed{\tilde{\Delta\theta}^{*}_{\text{out}} = \frac{1}{s}\left[\frac{1+K_1K_2}{K_2\sinh(K_1 s)\cosh(s)+\sinh(s)\cosh(K_1 s)} - \frac{1}{\sinh(s)}\right]} \quad (68)$$

Theoretically, K_1 and K_2 can be measured from an experimental thermal contrast curve through an "inverse" technique. The numerical inversion of the model is implemented by De Hoog's algorithm [10] whose MATLAB version (Invlap) is given in [11].

From K_1 and K_2 (or by a parameter substitution), it is also possible to calculate the thermal capacity and conductivity of the deposit by the following relations:

$$K_3 = K_1 K_2 = \frac{\rho_1 c_1 e_1}{\rho_2 c_2 e_2} \quad \text{thermal capacities ratio}$$

$$\text{or } K_3 = C_{t1}/C_{t2} \text{ with } C_{ti} = \rho c_i e_i \qquad \text{for } i = 1, 2 \quad (69)$$

and

$$K_4 = \frac{K_1}{K_2} = \frac{e_1}{e_2}\frac{\lambda_2}{\lambda_1} \quad \text{thermal resistances ratio}$$

$$\text{or } K_4 = R_1/R_2 \text{ with } R_i = e_i/\lambda_i \qquad \text{for } i = 1, 2 \quad (70)$$

Another parametrization of the same model consists in writing expression (68) as a function of K_3 and K_4.

The expression of the theoretical model with scaled parameters clearly shows that its output is in this case only function of two parameters. This means in particular that the thermophysical properties of the deposit can theoretically be obtained only if the properties of the substrate are known and as well as the thickness of each layer. Thus, the precision of the measurement also depends on the precision of these known parameters.

In the followings, our attention will be focused on two particular cases. The first one corresponds to a conductive deposit on an insulating material. The second one corresponds to an insulating film on a conductive substrate.

In these two cases, the materials we consider have low thicknesses and are good conductors. So, the Biot number based on the properties of the substrate $Bi = he_2/\lambda_2$ is low and it is possible, as a first approximation, to neglect its influence on the measured reduced rear face contrast ΔT^*.

It can be shown that even in the presence of heat losses, there is some kind of compensation through the construction of this contrast, which is a difference, which means that the present adiabatic model is a robust one: we will see in a later section that this parameter has a low influence in the estimation of the coating properties. The thicknesses and thermophysical properties are given in **Table 3**.

	Thickness (µm)	a (m^2/s)	λ (W/m/°K)	ρC_p (J/m3/°K)
Case 1 :	**Aluminium coating on a Cobalt/Nickel substrate**			
Film (1)	220	$9.46\ 10^{-5}$	230	$2.43\ 10^{6}$
Substrate (2)	1 100	$2.36\ 10^{-5}$	84.5	$3.57\ 10^{6}$
Case 2 :	**Insulating film on a Alumina substrate**			
Film (1)	247	$6.84\ 10^{-7}$	2.23	$3.26\ 10^{6}$
Substrate (2)	640	$7.47\ 10^{-6}$	23	$3.08\ 10^{6}$

Table 3: Thermophysical properties and thicknesses of the materials

The reduced thermograms for the substrate and two-layer material as well as the contrast curve are plotted for the conductive/insulating and insulating/conductive cases in **Figure 9** and **Figure 10** respectively.

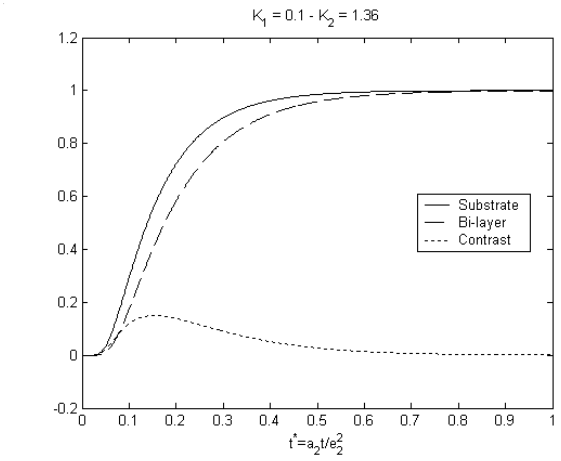


Figure 9 : *Case 1 – Conductive coating / Insulating substrate*

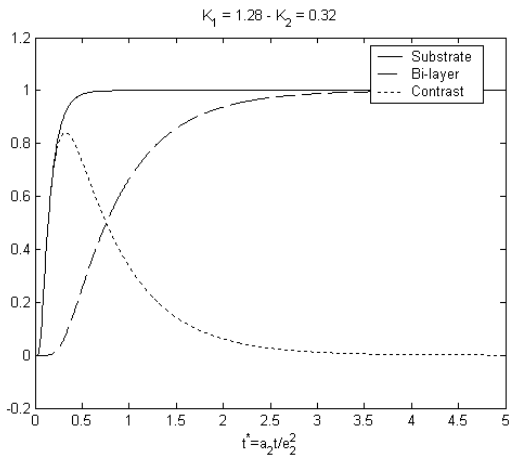


Figure 10 : *Case 2 – Insulating film / Conductive substrate*

- **Sensitivity Study**

The contrast curves and reduced sensitivities to parameters K_1 and K_2 for the two cases considered (conductive and insulating deposits) are plotted in **Figure 11** and **Figure 12**.

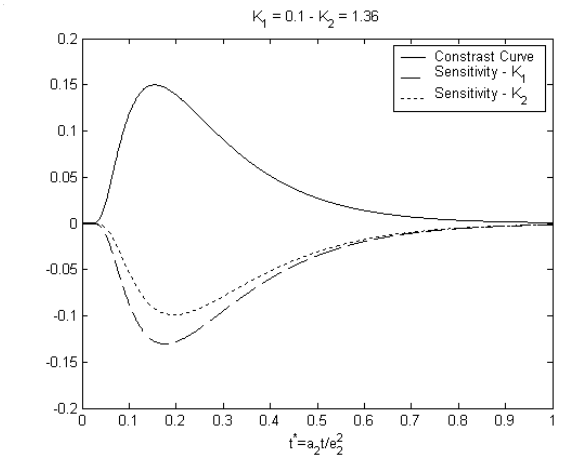


Figure 11 : *Contrast curve and reduced sensitivities to K_1 and K_2 (Case 1)*

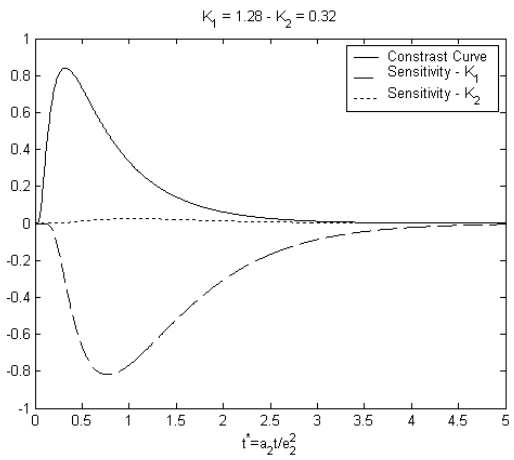


Figure 12 : *Contrast curve and reduced sensitivities to K_1 and K_2 (Case 2)*

These two examples are representative of most of the cases that can be met. In the first case, both sensibilities are of the same order of magnitude but seem to be strongly correlated: they exhibit a nearly constant ratio, which means that they are proportional. In the second case, one of the sensitivity is low.

- **Covariance and correlation matrices**

Table 4 gives the scaled covariance matrix $\mathrm{rcov}(\hat{\boldsymbol{K}}) = \sigma^2\left(\boldsymbol{S}^{*T}\boldsymbol{S}^{*}\right)^{-1}$ defined in (35), as well as the correlation matrix $\mathrm{cor}(\hat{\boldsymbol{K}})$ defined in (15), for the two cases considered (the standard-deviation of noise σ is taken equal to unity here and 1000 points in time are used for the simulation of the thermal contrast curve).

Scaled Variance-Covariance	**Scaled Variance-Covariance**
28.0302 -35.9846 -35.9846 **46.6417**	**0.1067** 3.1409 3.1409 **99.1677**
Correlation	**Correlation**
1.0000 -0.9952 -0.9952 1.0000	1.0000 0.9655 0.9655 1.0000
Case 1	**Case 2**

Table 4 : Reduced covariance and correlation matrices K_1 and K_2 (for $\sigma = 1$)

The most interesting information is given by the reduced variance-covariance matrix $\mathrm{rcov}(\hat{\boldsymbol{K}})$: it takes into account at the same times the reduced sensitivities through the inversion of the reduced information matrix $\boldsymbol{S}^{*T}\boldsymbol{S}^{*}$ as well as the noise through its standard deviation σ.

We calculate now the square root of the diagonal terms of matrix $\mathrm{rcov}(\hat{\boldsymbol{K}})$, that is the relative standard deviations of the estimates of each parameter K_1 and K_2, for a reduced standard deviation of the noise on each of the two T_2^* and $T_{1/2}^*$ scaled thermograms now equal to $\sigma^* = 0.01$. This corresponds to a signal over noise ratio of 100. So measurement of the (experimental) reduced thermal contrast ΔT^{*exp} is affected by a (relative) standard deviation ΔT^* equal to $\sqrt{2}\,\sigma^*$ (for two independent experiments, because $\mathrm{var}(\Delta T^{*exp}) = \mathrm{var}(T_2^{*\mathrm{exp}}) + \mathrm{var}(T_{1/2}^{*\mathrm{exp}}) = 2\,\sigma^{*2}$), one gets (application of equation (35) with $\sqrt{2}\,\sigma^*$ replacing σ):

- for case 1:
$$\begin{array}{lll} \sigma_{\hat{K}_1} / K_1 = \sqrt{2}\,\sigma^* \sqrt{28.0302} = 0.0749 \approx 7.5\% & \text{for} & K_1 = 0.1 \\ \sigma_{\hat{K}_2} / K_2 = \sqrt{2}\,\sigma^* \sqrt{46.6417} = 0.0966 \approx 9.5\,\% & \text{for} & K_2 = 1.36 \end{array} \tag{71}$$

It is interesting to calculate the singular values of the reduced sensitivity matrix $\boldsymbol{S}^*$ .They are the square roots of the eigenvalues (equal to the singular values) of the reduced information matrix $\boldsymbol{S}^{*T}\boldsymbol{S}^*$ and can also be calculated through the inverse of the eigenvalues of $\left(\boldsymbol{S}^{*T}\boldsymbol{S}^*\right)^{-1}$:

$$\begin{aligned} \mathrm{w}_1(\boldsymbol{S}^*) &= \left(\mathrm{w}_1(\boldsymbol{S}^{*T}\boldsymbol{S}^*)\right)^{1/2} = 1/\left(\mathrm{w}_2\left((\boldsymbol{S}^{*T}\boldsymbol{S}^*)^{-1}\right)\right)^{1/2} = 2.4347 \\ \mathrm{w}_2(\boldsymbol{S}^*) &= \left(\mathrm{w}_2(\boldsymbol{S}^{*T}\boldsymbol{S}^*)\right)^{1/2} = 1/\left(\mathrm{w}_1\left((\boldsymbol{S}^{*T}\boldsymbol{S}^*)^{-1}\right)\right)^{1/2} = 0.1159 \end{aligned} \tag{72}$$

This allows to get the condition number of $\boldsymbol{S}^*$ (see Lecture L3):

$$\mathrm{cond}(\boldsymbol{S}^*) = \mathrm{w}_1(\boldsymbol{S}^*)/\mathrm{w}_2(\boldsymbol{S}^*) = 21 \tag{73}$$

We can also calculate the root mean square reduced standard deviation m_q of the estimates of both parameters K_1 and K_2 defined in (37):

$$m_q = \sigma^* \sqrt{2}\,\left(1/w_1^2 + 1/w_2^2\right)^{1/2} = 0.0864 \tag{74}$$

It is easy to check that this value is simply the root mean square of the relative standards deviations given in (71).
Let us note that this value (73) is close to the lower bound of m_q defined in (38), here: $(\sigma^*\sqrt{2})/(\sqrt{2}\,w_2) = \sigma^*/w_2 = 0.0862$. The smallest singular value is mostly responsible for the relative errors on both parameters.

The same calculations can be made for the second case:

- for case 2:
$$\begin{array}{lll} \sigma_{\hat{K}_1} / K_1 = \sqrt{2}\,\sigma^* \sqrt{0.1067} = 0.0046 \approx 0.5\% & \text{for} & K_1 = 1.28 \\ \sigma_{\hat{K}_2} / K_2 = \sqrt{2}\,\sigma^* \sqrt{99.1677} = 0.1408 \approx 14.1\% & \text{for} & K_2 = 0.32 \end{array} \tag{75}$$

and : $$w_1(\boldsymbol{S}^*) = 11.7851 \qquad w_2(\boldsymbol{S}^*) = 0.1004 \tag{76}$$

So, the condition number of $\boldsymbol{S}^*$ is:

$$\text{cond}(\boldsymbol{S}^*) = w_1(\boldsymbol{S}^*)/w_2(\boldsymbol{S}^*) = 117 \tag{77}$$

which means that matrix $\boldsymbol{S}^*$ is more ill-conditioned in the second case with respect to the first one.

One also gets here:

$$m_q = 0.0996 \quad \text{and lower bound for } m_q : \sigma^* / w_2 = 0.0996 \tag{78}$$

So, returning to case 1, it appears clearly that both the ratios K_1 of the characteristic times and K_2 of the effusivities can be estimated with a relative error nearly equivalent for both parameters (in the 7 to 10 % interval): this was already apparent in **Figure 11** where the reduced sensitivity curves corresponding to both parameters were very close, with a slightly higher absolute value for the sensitivity to K_1.

For case 2, it is clearly the ratio K_1 of the characteristic times that can be reached, with a very good precision (0.5 % here): this is quite natural since the reduced sensitivity to K_2 in **Figure 12** is close to zero. So, because of the nonlinear character of this PEP problem, the accessible parameter depends on the location of the (K_1, K_2) parameter vector in the $\mathbf{R}^2$ plane. The question that remains is to know if is possible to measure, with higher precisions, two parameters derived from (K_1, K_2) using the experiment corresponding to case 1 for example. Let us introduce for instance the (K_3, K_4) pair instead of (K_1, K_2) in the analytical model.

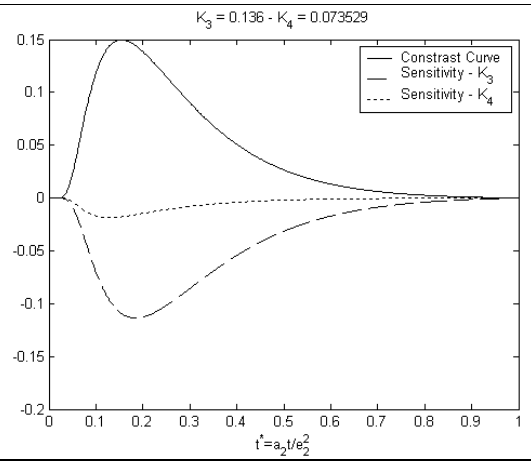


Figure 13 : *Contrast curve and reduced sensitivities to K_3 and K_4 - case 1*

Variance-Covariance

2.6921 -18.5189
-18.5189 **145.8475**

Correlation

1.0000 -0.9346
-0.9346 1.0000

Case 1

Table 5 : Reduced covariance and correlation matrices K_3 *and* K_4 (for σ = 1) - case 1

The thermal contrast is naturally the same (the materials are identical).

Table 5 gives the scaled covariance matrix $\text{rcov}(\hat{\boldsymbol{K}})$ as well as the correlation matrix $\text{cor}(\hat{\boldsymbol{K}})$ for the estimator of $\boldsymbol{K} = [K_3 \; K_4]^T$. The relative standard deviation of both parameters becomes (for $\sigma^* = 0.01$):

- for case 1:

$$\begin{aligned} &\sigma_{\hat{K}_3} / K_3 = \sqrt{2}\,\sigma^* \sqrt{2.6921} = 0.0232 \approx 2.3\% \quad \text{for} \quad K_3 = 0.136 \\ &\sigma_{\hat{K}_4} / K_4 = \sqrt{2}\,\sigma^* \sqrt{145.8475} = 0.1708 \approx 17.1\% \quad \text{for} \quad K_4 = 0.0735 \end{aligned} \tag{79}$$

So, when comparing (79) and (71), one clearly sees that instead of having (K_1 , K_2) with quite poor precisions, the (K_3, K_4) allows to retrieve very precise values for the ratio of volumetric heat capacities K_3. This was already apparent in **Figure 13**: the relative sensitivity to K_4 was quite low when compared to the one of K_3, but both minima of the corresponding curves occurred at times far apart, with a degree of collinearity much weaker than in figure 11 (see also section 3.3.2 of this lecture).

This result obtained for the two cases can be explained from the expression of the contrast curve.

$$\Delta\tilde{\theta}^* = \frac{1}{s}\left[\frac{1+K_1K_2}{K_2 \sinh(K_1 s)\cosh(s) + \sinh(s)\cosh(K_1 s)} - \frac{1}{\sinh(s)}\right] \tag{80}$$

In the previous case (conductive coating on an insulating substrate), K_1 is close to zero. A rough approximation can be obtained by setting: $\begin{cases} \sinh(K_1 s) \simeq K_1 s \\ \cosh(K_1 s) \simeq 1 \end{cases}$

$$\boxed{\Delta\tilde{\theta}^* = \frac{1}{s}\left[\frac{1+K_3}{K_3\, s \cosh(s) + \sinh(s)} - \frac{1}{\sinh(s)}\right]} \tag{81}$$

We can see then that within this first order approximation, the model is only a function of $K_3 = K_1K_2$. We can check the other criteria already considered for case 1 with the (K_1 , K_2) parameters:

$$w_1(\boldsymbol{S}^*) = 1.7270 \qquad w_2(\boldsymbol{S}^*) = 0.0821 \tag{82}$$

So, the condition number of $\boldsymbol{S}^*$ is:

$$\text{cond}(\boldsymbol{S}^*) = w_1(\boldsymbol{S}^*)/w_2(\boldsymbol{S}^*) = 21 \tag{83}$$

Compared to the preceding parameterization, the reduced sensitivity matrix $\boldsymbol{S}^*$ as well as its singular values have changed, but the condition number is the same, see (73).

One also gets here:

$$m_q = 0.1219 \quad \text{and lower bound for } m_q : \sigma^* / w_2 = 0.1218 \tag{84}$$

When both m_q's are compared, see (74), one can say that the global precision of the estimation of the (K_3, K_4) parameterization is lower than the (K_1, K_2) one. However, we will see later on that this superiority of the (K_3, K_4) parameterization is only an apparent one if both thermophysical characteristics of the film are looked for.

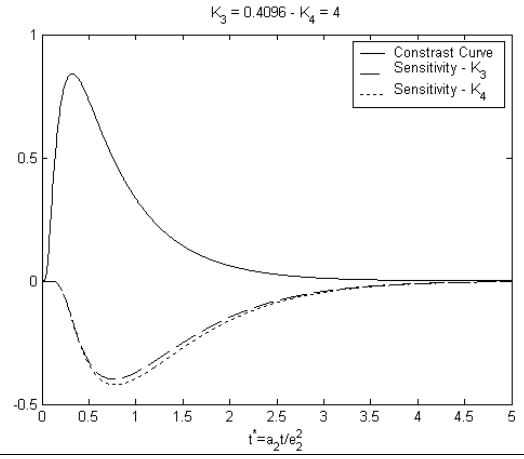


Figure 14 : *Contrast curve and sensitivities to K_3 and K_4 (Case 2)*

Variance-Covariance
103.5845 -97.1801 -97.1801 **91.1985**
Correlation
1.0000 -0.9999 -0.9999 1.0000
Case 2

Table 6 : Reduced covariance and correlation matrices K_3 *and* K_4 (for σ = 1)

In case 2 (insulating coating on a conductive substrate), parameters K_3 and K_4 are strongly correlated and exhibit the same sensitivity curves – see **Figure 14**. This confirms the result we observed previously, that is a thermal contrast mostly sensitive to K_1.

$$K_3\ K_4 = \frac{C_1}{C_2}\frac{R_1}{R_2} = \frac{R_1 C_1}{R_2 C_2} = \frac{tc_1}{tc_2} = {K_1}^2 \tag{85}$$

This can be also explained by the fact that K_1 is close to unity:

$$\sinh(K_1 s)\cosh(s) \simeq K_1 \sinh(s)\cosh(K_1 s) \tag{86}$$

This yield:

$$\Delta\tilde{\theta}^*_{out} = \frac{1}{s}\left[\frac{1}{\sinh(s)\cosh(K_1 s)} - \frac{1}{\sinh(s)}\right] \tag{87}$$

So, the thermal contrast is mainly a function of K_1. Returning to the same calculation as in the other case, using **Table 6**, one gets:

- for case 2:
$$\boxed{\begin{array}{lll} \sigma_{\hat{K}_3} / K_3 = \sqrt{2}\,\sigma^* \sqrt{103.5845} = 0.1439 \approx 14.4\% & \text{for} & K_3 = 0.4096 \\ \sigma_{\hat{K}_4} / K_4 = \sqrt{2}\,\sigma^* \sqrt{91.1985} = 0.1351 \approx 13.5\,\% & \text{for} & K_4 = 4 \end{array}} \quad (88)$$

The singular values of the reduced sensitivity matrix are:

$$w_1(\boldsymbol{S}^*) = 8.3624 \qquad w_2(\boldsymbol{S}^*) = 0.0717 \quad (89)$$

So, the condition number of $\boldsymbol{S}^*$ is:

$$\text{cond}(\boldsymbol{S}^*) = w_1(\boldsymbol{S}^*)/w_2(\boldsymbol{S}^*) = 117 \quad (90)$$

We observe here the same thing as for case 2: the condition number of the reduced sensitivity matrix is independent of the parameterization, see (77).

One also gets here:

$$m_q = 0.1396 \quad \text{and lower bound for } m_q : \sigma^* / w_2 = 0.1395 \quad (91)$$

When both m_q's are compared, see (78), one can say that the global precision of the estimation of the (K_3, K_4) parameterization, which provided an excellent estimation for K_3. is lower than the (K_1, K_2) one.

4.3 Note on the change of parameters

It has been suggested earlier that some change of parameterization would allow to overcome parameter estimation difficulties such as in the case of high correlation coefficients inducing high variances for the estimated parameters for example. We want here to come back to this discussion to give, very briefly, some precisions and our conclusions.

First, and taking experience of what has been shown previously, if a change of parameterization is made that results in the production of a new parameter of sensitivity close to zero (and thereof excluded from the model), this new parameterization will have a positive effect and will allow to properly estimate the remaining ones. Note that it is the object of Dimensional Analysis to help making such reparameterization efficient.

Second, if all the parameters of the problem have non negligible sensitivities but appear correlated, the question is: is it possible to find a new set of parameters defined from the initial one, to enhance the quality of the estimation process?

The answer is no. It can be demonstrated, see Remy [9] that the sensitivities to a new set of parameters can be derived from the sensitivities of the current set (using the Jacobian of the transformation). The same is true for the variance-covariance matrix and the explanation is obvious from the quantified SVD analysis given above (the same condition number of $\boldsymbol{S}^*$ is obtained whatever set of parameterizations is used) These relationships show that:

- if two parameters appear correlated in a given set of parameters, two parameters of a new set, recombined from the previous ones, will also be correlated,

- if the sensitivity of a parameter is changed with a new parameterization (for example, it is enhanced), this will not change its variance ultimately.

For instance, if we keep the parameter K_1 and choose another second parameter instead of K_2, we can show that the sensitivity curve to K_1 can become higher or lower: we have to remind that the partial derivative that appears in the definition (4) of a sensitivity coefficient is associated to the variation of the output of the model for a variation of a given parameter, which requires that the other ones stay fixed at given values. This means that if the definition of these other parameters is changed, such is also the case for the sensitivity coefficients. So, talking of a sensitivity coefficient to a given parameter does not mean anything if the other parameters in the parameter vector are not specified.

So, one can wonder whether it would be possible to improve the estimation of K_1 by combining this parameter with a particular parameter that can increase its sensitivity. In fact, this is not true because the standard-deviations of the estimates of the new parameters do not only depend on the sensitivities of the old parameters but also on the correlation between the estimates of the old parameters.

To show this, we are going to see through an example how the standard-deviations (square roots of variances) of the new set of parameters change when one parameter is kept as for instance parameter K_1, that is $K_a = K_1^{\alpha} K_2^{\beta}$ with α=1; β=0, while K_2 is replaced by $K_b = F_b(K_1, K_2)$:

$$\begin{aligned} K_a &= F_a(K_1) = K_1 \\ K_b &= F_b(K_1, K_2) \end{aligned} \tag{92}$$

We have:

$$\boldsymbol{y}_{mo} = \boldsymbol{\eta}(\boldsymbol{t}; \boldsymbol{K}) \text{ with } \boldsymbol{K} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \Rightarrow \mathrm{d}\boldsymbol{y}_{mo} = \boldsymbol{S}\ \mathrm{d}\boldsymbol{K} = \boldsymbol{S}'\ \mathrm{d}\boldsymbol{K}' \text{ with } \boldsymbol{K}' = \begin{bmatrix} K_a \\ K_b \end{bmatrix} \tag{93}$$

where $\boldsymbol{S}$ is the sensitivity matrix to the old (K_1, K_2) set of parameters and $\boldsymbol{S}$' the sensitivity matrix to the new (K_a, K_b) one. This requires the calculation of the Jacobian matrix $\boldsymbol{J}$ of this transformation since;

$$\mathrm{d}\boldsymbol{K}' = \boldsymbol{J}\ \mathrm{d}\boldsymbol{K} \quad \Rightarrow \quad \boldsymbol{S} = \boldsymbol{S}'\boldsymbol{J} \qquad \text{and} \qquad \mathrm{cov}(\hat{\boldsymbol{K}}') = \boldsymbol{J}\ \mathrm{cov}(\hat{\boldsymbol{K}})\ \boldsymbol{J}^T \tag{94}$$

The last equation in (94) stems from the linearization around the exact value of the $\boldsymbol{K}$ parameter vector:

$$\mathrm{cov}(\hat{\boldsymbol{K}}) = \mathrm{cov}(\mathrm{d}\hat{\boldsymbol{K}}) \tag{95}$$

with:
$$\boldsymbol{J} = \frac{D(F_a, F_b)}{D(K_1, K_2)} = \begin{bmatrix} \frac{\partial F_a}{\partial K_1} & \frac{\partial F_a}{\partial K_2} \\ \frac{\partial F_b}{\partial K_1} & \frac{\partial F_b}{\partial K_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ F_{b,1} & F_{b,2} \end{bmatrix} \quad (96)$$

So, the sensitivity matrix to the new parameter set ***K'*** is:

$$\boldsymbol{S}' = [\boldsymbol{S}_a \quad \boldsymbol{S}_b] = \boldsymbol{S}\boldsymbol{J}^{-1} = [\boldsymbol{S}_1 \quad \boldsymbol{S}_2] \begin{bmatrix} 1 & 0 \\ -F_{b,1}/F_{b,2} & 1/F_{b,2} \end{bmatrix} = [\boldsymbol{S}_1 - (F_{b,1}/F_{b,2})\boldsymbol{S}_2 \quad (1/F_{b,2})\boldsymbol{S}_2] \quad (97)$$

Here the old sensitivity column vectors $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, as well as the new ones $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$, have been explicitly written in terms of the corresponding sensitivity matrices, $\boldsymbol{S}$ and $\boldsymbol{S}'$ respectively.

Application of (94) allows the calculation of the variances and covariance of the estimators of the new set of parameters (K_a, K_b):

$$\mathrm{cov}(\hat{\boldsymbol{K}}') = \begin{bmatrix} \mathrm{var}(\hat{K}_a) & \mathrm{cov}(\hat{K}_a, \hat{K}_b) \\ \mathrm{cov}(\hat{K}_a, \hat{K}_b) & \mathrm{var}(\hat{K}_b) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ F_{b,1} & F_{b,2} \end{bmatrix} \begin{bmatrix} \mathrm{var}(\hat{K}_1) & \mathrm{cov}(\hat{K}_1, \hat{K}_2) \\ \mathrm{cov}(\hat{K}_1, \hat{K}_2) & \mathrm{var}(\hat{K}_2) \end{bmatrix} \begin{bmatrix} 1 & F_{b,1} \\ 0 & F_{b,2} \end{bmatrix} \quad (98)$$

that is:

$$\begin{aligned} &\mathrm{var}(\hat{K}_a) = \mathrm{var}(\hat{K}_1) \\ &\mathrm{var}(\hat{K}_b) = F_{b,1}{}^2 \mathrm{var}(\hat{K}_1) + F_{b,2}{}^2 \mathrm{var}(\hat{K}_2) + 2F_{b,1}F_{b,2}\,\mathrm{cov}(\hat{K}_1, \hat{K}_2) \\ &\mathrm{cov}(\hat{K}_a, K_b) = F_{b,1}\,\mathrm{var}(\hat{K}_1) + F_{b,2}\,\mathrm{cov}(\hat{K}_1, \hat{K}_2) \end{aligned} \quad (99)$$

We can see that even if the change of parameters modifies the sensitivity to parameter K_a , that replaces parameter K_1 in the new set of parameters, the variance of this parameter remains unchanged whatever the choice of the second parameter.

This means that the variance of a given parameter (and consequently the error on this parameter) is independent on the choice of the second parameter. Thus, identifying the parameter K_1 from the (K_1, K_2) pair is equivalent to estimating K_1 from the (K_1, K_3) or (K_1, K_4) pairs.

Similarly, we can show that estimating parameters (K_3, K_4) either through the parameterization (K_1, K_2) or directly, is strictly the same.

The conclusion is that the interest of a change of parameters is justified only when an improved estimation of a particular parameter of interest is looked for.

Whatever the parameterization, if the thicknesses of both layers are known, as well as the thermophysical properties of the substrate, we have:

$$\begin{aligned} \sigma_{\rho\hat{c}_1} / \rho c_1 &= \sigma_{\hat{K}_3} / K_3 = 2.3\ \% \quad \text{for case 1} \\ \sigma_{\hat{a}_1} / a_1 &= \sigma_{\hat{K}_1} / K_1 = 0.5\ \% \quad \text{for case 2} \end{aligned} \tag{100}$$

These relative standard deviation of the estimated thermophysical properties of the front face layer are valid for a signal to noise ratio equal to 100 for the experimental thermogram of each flash experiment (single substrate layer and two-layer sample). So, this rear face thermal constrast technique allows estimation of the capacity of the film for case 1 and of its diffusivity in case 2, for high enough signal over noise ratios.

In case of very low sensitivity to a given parameter, it is possible to fix the value of the corresponding parameter to its nominal values. So, if the number of parameters that are looked for is reduced, then the stochastic errors on the remaining parameters (reduced standard deviations) decrease. However, their estimation becomes biased and leads to a systematic error on each estimated parameter such as:

$$\boldsymbol{b}_{\hat{\boldsymbol{\beta}}_r} = \mathrm{E}\left(\hat{\boldsymbol{\beta}}_r\right) - \boldsymbol{\beta}_r = -\left(\boldsymbol{S}_r^T \boldsymbol{S}_r\right)^{-1} \boldsymbol{S}_r^T \boldsymbol{S}_c \left(\boldsymbol{\beta}_c^{nom} - \boldsymbol{\beta}_c^{exact}\right) \tag{101}$$

Here the initial parameter vector has been decomposed into two parts $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_c \end{bmatrix}$, where $\boldsymbol{\beta}_r$ gathers the parameters that are looked for and its complementary part $\boldsymbol{\beta}_c$ is supposed to be known, that is its value is blocked to a nominal value $\boldsymbol{\beta}_c = \boldsymbol{\beta}_c^{nom}$ which differs from its exact value $\boldsymbol{\beta}_c^{exact}$. Equation (101), which has already been derived in the case of a linear model in lecture L3 of this series (see also [1]), corresponds here to a linearization in the neighborhood of the exact value of $\boldsymbol{\beta}$.

This technique, which consists in reducing the number of parameters that are looked for, presents an interest only if the bias caused by the reduction of the number of parameters and its associated standard deviations are much lower than the initial stochastic error as illustrated in **Figure 15**.

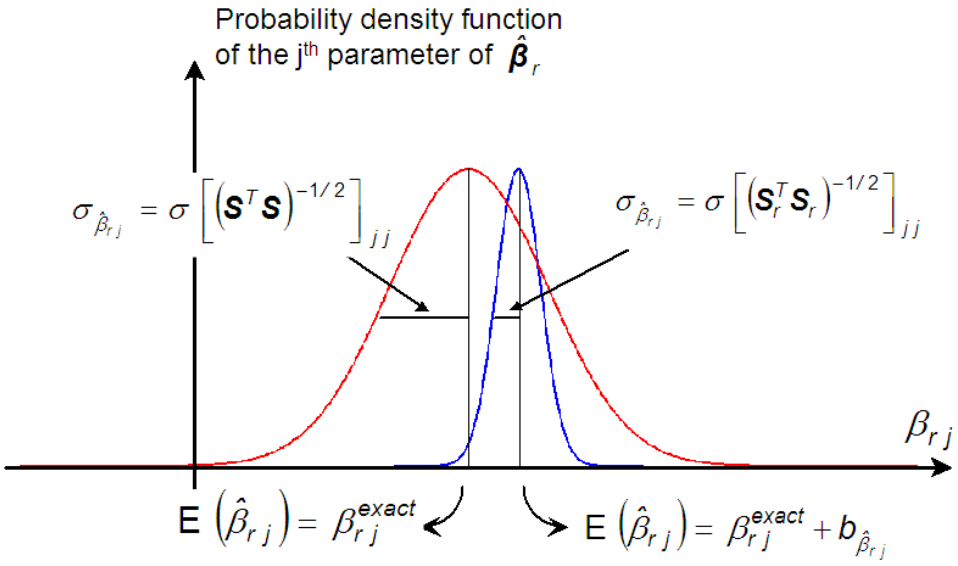


Figure 15 : *Comparison between the probability density distributions of the j^{th} parameter of the parameter vector for two different estimators 1) all the parameters in $\boldsymbol{\beta}$ are estimated altogether (red) or 2) only the components of one of its part $\boldsymbol{\beta}_r$ (blue) are estimated while its complementary part $\boldsymbol{\beta}_c$ are blocked to its nominal value.* ***NB:*** *here one assumes that index j in $\boldsymbol{\beta}$ and in $\boldsymbol{\beta}_r$ are the same ($\beta_{rj} = \beta_j$) and that the scale of the vertical axis is different for both distributions for practical plotting reasons (the area below both distributions should be equal to unity)*

5. Conclusion

Useful tools have been introduced for the analysis of estimations (variance-covariance matrix) and the detection of the ill-conditioned character of the Parameter Estimation Problem (PEP). Different techniques have been presented for tracking the true degrees of freedom of a given PEP (matrix rank, correlations between parameters, SVD, ...). If we want to enhance the estimation of a given parameter, one solution is to use a reduced model. This reduced model can be either unbiased or biased. It is of particular interest to know if a reduced model is biased or not.

We have proposed, in the last section of the lecture, to work with a variable estimation time interval in order to evaluate the systematic error caused in the estimated parameters. We hope that the different "realistic" examples of thermal metrology presented in this lecture will help the reader to master the corresponding tools to get good estimates in a PEP.

References

[1] Chapter 9: Nonlinear estimation techniques, B. Rémy, S. André, in *Thermal Measurements and Inverse Techniques,* 2011, Orlande H R.B., Fudym O., Maillet D. and Cotta R. M., editors: CRC Press, Series: Heat Transfer, 770p.

[2] J.V. Beck, K..J. Arnold, Parameter Estimation in Engineering and Science, John Wiley & Sons, 1977.

[3] K. Levenberg, A method for the solution of certain problems in Least Squares, Quart. Appl. Math. 2, 164-168, 1944.

[4] Gallant, A.R., 1975, Nonlinear regression, *Am. Stat.*, 29:73-81.

[5] Andre S., Serra J.J., Cella N and Silva Neto A.J., Parameter Estimation of Thermo-optical Glass Properties From Experimental Phase Lag Signals Obtained Through Periodic Heat Flux Excitation, *Proc. 4th International Conference on Inverse Problems in Engineering: Theory and Practice*, May 26-31, 2002, Angra dos Reis, Brasil.

[6] The Physical Basis of dimensional analysis, Ain A. Sonin, MIT course manuscript, 2001 http://me.mit.edu/people/*sonin*/html

[7] Maillet D., André S., Batsale J.C., Degiovanni A. and Moyne C., Thermal Quadrupoles : Solving the Heat equation Through Integral Transforms, Chichester, PA: John Wiley & Sons Ltd, 2000.

[8] Chapter 1: Modelling in heat transfer, J.-L. Battaglia and D. Maillet, in Thermal Measurements and Inverse Techniques, 2011, Orlande H.R.B., Fudym O., Maillet D. and Cotta R. M., editors: CRC Press, Series: Heat Transfer, 770p.

[9] B. Remy and A. Degiovanni, "Parameters Estimation and Measurement of Thermophysical Properties of Liquids", International Journal of Heat & Mass Transfer – Vol. 48,Issues 19-20, September 2005, pp 4103-4120.

[10]. F.R. de Hoog, J.H. Knight, A.N. Stokes, An improved method for numerical inversion of Laplace transforms, SIAM. J. Sci. Stat. Comp. 3, 357-366, (1982)

[11] http://www.mathworks.com/matlabcentral/answers/uploaded_files/1034/invlap.m

Appendix 1 - Reminder of the Singular Value Decomposition of a rectangular matrix

Any rectangular matrix (called $\boldsymbol{K}$ here) with real coefficients and of dimensions (m, n) with $m \geq n$, can be written under the form :

$$\boldsymbol{K} = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^T \text{, that is } \begin{bmatrix} & \\ & \boldsymbol{K} \\ & \end{bmatrix} = \begin{bmatrix} & \\ & \boldsymbol{U} \\ & \end{bmatrix} \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & w_n \end{bmatrix} \begin{bmatrix} & \\ & \boldsymbol{V}^T \\ & \end{bmatrix} \quad \text{(A1a)}$$

This expression is sometimes called "lean" singular decomposition or "economical" SVD and involves

- $\boldsymbol{U}$, an orthogonal matrix of dimensions (m, n), its column vectors (the *left* singular vectors of $\boldsymbol{K}$) have a unit norm and are orthogonal by pairs: $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the identity matrix of dimension n. Its columns are composed of the first n eigenvectors $\boldsymbol{U}_k$, ordered according to decreasing values of the eigenvalues of matrix $\boldsymbol{K}\,\boldsymbol{K}^T$. Let us note that, in the general case, $\boldsymbol{U}\boldsymbol{U}^T \neq \boldsymbol{I}_m$.

- $\boldsymbol{V}$, a square orthogonal matrix of dimensions (n, n), $\boldsymbol{V}\,\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n$. Its column vectors (the *right* singular vectors of $\boldsymbol{K}$), are the n eigenvectors $\boldsymbol{V}_k$, ordered according to decreasing eigenvalues, of matrix $\boldsymbol{K}^T\boldsymbol{K}$;

- $\boldsymbol{W}$, a square diagonal matrix of dimensions (n, n), that contains the n so-called *singular* values of matrix $\boldsymbol{K}$, ordered according to decreasing values: $w_1 \geq w_2 \geq \cdots \geq w_n$. The singular values of matrix $\boldsymbol{K}$ are defined as the square roots of the eigenvalues of matrix $\boldsymbol{K}^T\boldsymbol{K}$. If matrix $\boldsymbol{K}$ is square and symmetric, the eigenvalues and the singular values of $\boldsymbol{K}$ are the same.

Another SVD form called "Full Singular Value Decomposition" is available for matrix $\boldsymbol{K}$. In this equivalent definition, both matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ are changed: the matrix replacing $\boldsymbol{U}$ is now square (size m x m) and the matrix replacing $\boldsymbol{W}$ is now diagonal but non square (size m x n). In the case $m \geq n$, this can be written:

$$\boldsymbol{K} = \boldsymbol{U}_0\,\boldsymbol{W}_0\,\boldsymbol{V}^t \quad \text{with} \quad \boldsymbol{U}_0 = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U}_{comp} \end{bmatrix};\ \boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{W} \\ \mathbf{0}_{(m-n)\ \mathrm{x}\ n} \end{bmatrix} \quad \text{and} \quad \dim(\boldsymbol{U}_{comp}) = m\ x\ (m - n) \quad \text{(A1b)}$$

or:

$$\begin{bmatrix} \\ \boldsymbol{K} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \boldsymbol{U} \qquad \boldsymbol{U}_{comp} \\ \\ \end{bmatrix} \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & w_n \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \\ \boldsymbol{V}^t \\ \\ \end{bmatrix} \qquad \text{(A1c)}$$

Matrix $\boldsymbol{U}_{comp}$ is composed of the (m - n) left singular column vectors do not present in $\boldsymbol{U}$. So, the concatenated matrix $\boldsymbol{U}_0$ verifies now:

$$\boldsymbol{U}_0^t \boldsymbol{U}_0 = \boldsymbol{U}_0 \boldsymbol{U}_0^t = \boldsymbol{U}\boldsymbol{U}^t + \boldsymbol{U}_{comp}\boldsymbol{U}_{comp}^t = \boldsymbol{I}_m \qquad \text{(A1d)}$$

This singular value decomposition (A1b) can be implemented for any matrix $\boldsymbol{K}$, with real value coefficients, for $m \geq n$.

Appendix 2 - Singular Value Decomposition of the scaled sensitivity matrix

This singular value decomposition can be implemented for any matrix $\boldsymbol{K}$.

A double change of basis, in the measurement domain and in the parameter domain, using the matrices of the left $\boldsymbol{U}$ and right $\boldsymbol{V}$, in the SVD of $\boldsymbol{S}^*$ written for $\boldsymbol{K} = \boldsymbol{S}^*$ yields:

$$\boldsymbol{S}^* = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^T \qquad \text{(A2)}$$

Matrix $\boldsymbol{V}$ is used as a (square) change of matrix basis and it transforms the differential of the reduced parameter vector $\mathrm{d}\boldsymbol{x}$, see (29) into a new differential vector $\mathrm{d}\boldsymbol{p}$, where $\boldsymbol{p}$ can be called the diagonal parameter vector, of dimensions (n, 1).

Matrix $\boldsymbol{U}$ allows to change the differential observation vector $\mathrm{d}\boldsymbol{y}_{mo}$ of dimensions (m, 1) into a differential vector $\mathrm{d}\boldsymbol{z}_{mo}$ of smaller length, where $\boldsymbol{z}_{mo}$ can be called the diagonal observation vector, of dimensions (n, 1).

$$\mathrm{d}\boldsymbol{y}_{mo} = \boldsymbol{U}\ \mathrm{d}\boldsymbol{z}_{mo} \qquad \text{and} \qquad \mathrm{d}\boldsymbol{x} = \boldsymbol{V}\ \mathrm{d}\boldsymbol{p} \qquad \text{(A3a, b)}$$

Let us note here that the reduction of the length of the observation vector (m observations for $\mathrm{d}\boldsymbol{y}_{mo}$ and only n components in $\mathrm{d}\boldsymbol{z}_{mo}$ stems from the fact that the (m-n) singular eigenvectors $\boldsymbol{U}_k$ not present in matrix $\boldsymbol{U}$ corresponds to null singular values w_k (for $k > n$).

Use of equations (A1) to (A3), together with the property $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n$, allows to get the equivalent of the differential model (31a) in the double transformed space:

$$\mathrm{d}\boldsymbol{z}_{mo} = \boldsymbol{W}\ \mathrm{d}\boldsymbol{p} \qquad \text{(A4)}$$

This equation corresponds to a diagonalization of the model in R^n, and one gets then, component by component:

$$dp_k = \frac{1}{w_k} dz_{mo,k} \qquad for \qquad k = 1, 2, \cdots, n \tag{A5}$$

Combining (A3a, b) and (A4) yields:

$$d\boldsymbol{x} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{U}^T d\boldsymbol{y}_{mo} = \boldsymbol{S}^{*+} d\boldsymbol{y}_{mo} \tag{A6}$$

where $\boldsymbol{S}^{*+} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{U}^T$ is the pseudo-inverse, or Moore-Penrose inverse, of the scaled sensitivity matrix $\boldsymbol{S}^*$.

Combination of the preceding equations leads to a relationship between $d\boldsymbol{\beta}$ and $d\boldsymbol{y}_{mo}$:

$$d\boldsymbol{\beta} = \boldsymbol{R}_{nom}\boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T d\boldsymbol{y}_{mo} \tag{A7}$$

and an integration can be implemented to give the relationship between the diagonal and original sets of parameters in a column vector form:

$$\boldsymbol{p} = \boldsymbol{V}^T\boldsymbol{x} = \boldsymbol{V}^T \ln\left(\boldsymbol{R}_{nom}^{-1}\boldsymbol{\beta}\right) \approx \boldsymbol{V}^T\boldsymbol{R}_{nom}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{nom}\right) \quad \text{because} \quad \boldsymbol{p}^{nom} = \boldsymbol{V}^T\mathbf{x}^{nom} = \mathbf{0} \tag{A8}$$

The transformed observation vector can be expressed:

$$\boldsymbol{z}_{mo} = \boldsymbol{U}^T(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})) = \boldsymbol{W}\,\boldsymbol{p} \qquad \text{because} \qquad \boldsymbol{z}_{mo}^{nom} = \boldsymbol{W}\,\boldsymbol{p}^{nom} = \mathbf{0} \tag{A9}$$

Combining (A8) and (A9) yields:

$$\boldsymbol{p} = \boldsymbol{V}^T \ln\left(\boldsymbol{R}_{nom}^{-1}\boldsymbol{\beta}\right) = \boldsymbol{W}^{-1}\boldsymbol{U}^T\left(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) \Rightarrow \boldsymbol{\beta} = \boldsymbol{R}_{nom} \exp\left(\boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T\left(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right)\right) \tag{A10}$$

An approximation of this expression in the neighbourhood of $\boldsymbol{\beta}^{nom}$ is available:

$$\boldsymbol{\beta} \approx \boldsymbol{R}_{nom}\left[\mathbf{1} + \boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T\left(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right)\right] = \boldsymbol{\beta}^{nom} + \boldsymbol{R}_{nom}\boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T\left(\boldsymbol{y}_{mo} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) \tag{A11}$$

where **1** is the column vector of length *n* whose coefficients are equal to unity.

Appendix 3 – Non-linear Ordinary Least Square estimator and SVD

It is interesting to compare diagonal equation (A5) that shows the interest of an inversion in the left and right singular spaces with the OLS estimator (12) of parameter $\boldsymbol{\beta}$. So, if the first order approximation in the neighbourhood of $\boldsymbol{\beta}^{nom}$ is considered, the difference between

measurements and model outputs can be expressed with the *residual* vector defined in (10), and $\boldsymbol{r}_{lin}$ the linearized form of this difference vector:

$$\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}) \approx \boldsymbol{r}_{lin}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom}) - \boldsymbol{S}(\boldsymbol{\beta}^{nom})(\boldsymbol{\beta} - \boldsymbol{\beta}^{nom}) \quad \text{(A12)}$$

The least squares sum J_{OLS} can be written as a quadratic form $J^{\diamond}$, using the fact that $J_{OLS} = J_{OLS}{}^{T}$ (scalar):

$$J(\boldsymbol{\beta}) = \boldsymbol{r}^T(\boldsymbol{\beta})\,\boldsymbol{r}(\boldsymbol{\beta}) \approx J^{\diamond}(\boldsymbol{\beta}) = \left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right)^T \left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{nom})^T \boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\boldsymbol{S}(\boldsymbol{\beta}^{nom})(\boldsymbol{\beta} - \boldsymbol{\beta}^{nom}) - 2(\boldsymbol{\beta} - \boldsymbol{\beta}^{nom})^T \boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) \quad \text{(A13)}$$

When the minimum is reached, one gets:

$$\frac{\mathrm{d}J^{\diamond}}{\mathrm{d}\boldsymbol{\beta}} = 0 \qquad \Rightarrow \boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\,\boldsymbol{S}(\boldsymbol{\beta}^{nom})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{nom}) = \boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) \quad \text{(A14)}$$

which leads to an approximation of the OLS estimator:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{nom} = \left(\boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\boldsymbol{S}(\boldsymbol{\beta}^{nom})\right)^{-1}\boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) \quad \text{(A15)}$$

This is exactly the same equation as the iterative algorithm (12), with $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}^{(k+1)}$ and $\boldsymbol{\beta}^{nom} = \hat{\boldsymbol{\beta}}_{OLS}^{(k)}$. One shows, using (31b) and (A2):

$$\left(\boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\boldsymbol{S}(\boldsymbol{\beta}^{nom})\right)^{-1}\boldsymbol{S}^T(\boldsymbol{\beta}^{nom}) = \boldsymbol{R}_{nom}\,\boldsymbol{V}\,\boldsymbol{W}^{-1}\,\boldsymbol{U}^T \quad \text{(A16)}$$

The least square estimator (A15), with the diagonal parameter $\boldsymbol{p}$ and the experimental diagonal signal $\boldsymbol{z}$ in their new bases, can be written thanks to (A16):

$$\hat{\boldsymbol{p}} = \boldsymbol{W}^{-1}\boldsymbol{z} \quad \text{with} \quad \boldsymbol{z} = \boldsymbol{U}^T(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})) \quad \text{(A17a, b)}$$

Equation (A17a) is diagonal. Use of (A15) and (A16) provides a new expression for the OLS estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{nom}\left(\boldsymbol{1} + \boldsymbol{V}\,\boldsymbol{W}^{-1}\boldsymbol{U}^T(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom}))\right) \quad \text{(A18)}$$

This expression is the same as relationship (A1) that links $\boldsymbol{\beta}$ and $\boldsymbol{y}_{mo}(\boldsymbol{\beta})$: these corresponding two values are simply replaced by the linearized OLS estimator $\hat{\boldsymbol{\beta}}$ and by measurements $\boldsymbol{y}$ respectively.

The linearized OLS estimator of the reduced parameter vector stems directly from (A19):

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{S}^{*T}(\boldsymbol{\beta}^{nom})\,\boldsymbol{S}^{*}(\boldsymbol{\beta}^{nom})\right)^{-1}\boldsymbol{S}^{*T}(\boldsymbol{\beta}^{nom})\left(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{nom})\right) \quad \text{(A20)}$$

Appendix 4 – Variance-covariance of the Non-linear Ordinary Least Square estimator and SVD

With the noise properties defined in (8), the variance-covariance of the linearized OLS estimator $\hat{\boldsymbol{\beta}}$ given by equation (A15), can be written thanks to (31b) and (A2):

$$\begin{aligned}\mathrm{cov}(\hat{\boldsymbol{\beta}}) &= \sigma^2 \left(\boldsymbol{S}^T(\boldsymbol{\beta}^{nom})\boldsymbol{S}(\boldsymbol{\beta}^{nom})\right)^{-1} = \sigma^2 \left(\boldsymbol{R}_{nom}^{-1}\boldsymbol{S}^{*T}\boldsymbol{S}^{*}\boldsymbol{R}_{nom}^{-1}\right)^{-1} \\ &= \sigma^2 \boldsymbol{R}_{nom}\left(\boldsymbol{S}^{*T}\boldsymbol{S}^{*}\right)^{-1}\boldsymbol{R}_{nom} = \sigma^2 \boldsymbol{R}_{nom}\boldsymbol{V}\boldsymbol{W}^{-2}\boldsymbol{V}^T\boldsymbol{R}_{nom}\end{aligned} \tag{A21}$$

This expression is valid if the difference between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{nom}$ is small: it is always the case near convergence of algorithm (12) where $\boldsymbol{\beta}^{nom}$ can be redefined as $\boldsymbol{\beta}^{nom} = \hat{\boldsymbol{\beta}}_{OLS}^{(k)}$ and with $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}^{(k+1)}$.

The expression of the variance-covariance matrix of $\hat{\boldsymbol{x}} = \boldsymbol{R}_{nom}^{-1}\hat{\boldsymbol{\beta}}$ becomes:

$$\mathrm{cov}(\hat{\boldsymbol{x}}) = \boldsymbol{R}_{nom}^{-1}\mathrm{cov}(\hat{\boldsymbol{\beta}})\left(\boldsymbol{R}_{nom}^{-1}\right)^T = \sigma^2\,\boldsymbol{V}\,\boldsymbol{W}^{-2}\boldsymbol{V}^T \tag{A22a}$$

The first relationship in equation (A22a) allows to calculate the reduced covariance matrix of $\hat{\boldsymbol{\beta}}$, $\mathrm{rcov}(\hat{\boldsymbol{\beta}})$, whose diagonal coefficients are the reduced variances of the estimators of each parameter, using the nominal values of the parameters as scaling factors:

$$\mathrm{rcov}(\hat{\boldsymbol{\beta}}) = \mathrm{cov}(\hat{\boldsymbol{x}}) = \begin{bmatrix} \sigma_{\hat{\beta}_1}^2/(\beta_1^{nom})^2 & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_2)/(\beta_1^{nom}\beta_2^{nom}) & \cdots & \mathrm{cov}(\hat{\beta}_1,\hat{\beta}_n)/(\beta_1^{nom}\beta_n^{nom}) \\ & \sigma_{\hat{\beta}_2}^2/(\beta_2^{nom})^2 & & \\ & & \ddots & \\ \text{Symmetric} & & & \sigma_{\hat{\beta}_n}^2/(\beta_n^{nom})^2 \end{bmatrix} = \sigma^2\left(\boldsymbol{S}^{*T}\boldsymbol{S}^{*}\right)^{-1} \tag{A22b}$$

where $\sigma_{\hat{\beta}_j}$ is the standard deviation of $\hat{\beta}_j$. The square roots of the diagonal terms of this matrix, $\sigma_{\hat{\beta}_1}/\beta_1^{nom}$, can be considered as a measure of the relative error made for each parameter and caused by presence of noise in the measurements $\boldsymbol{y}$.

It is very interesting to calculate the trace of this matrix, which is equal to the sum of the variances of the different components of $\hat{\boldsymbol{x}}$:

$$\begin{aligned}&\mathrm{Tr}\left(\mathrm{cov}(\hat{\boldsymbol{x}})\right) \equiv \sum_{j=1}^{n}\sigma_{xj}^2 = \sum_{j=1}^{n}\left(\sigma_{\beta j}/\beta_j^{nom}\right)^2 \\ \Rightarrow\quad &\sum_{j=1}^{n}\left(\sigma_{\beta j}/\beta_j^{nom}\right)^2 = \sigma^2\,\mathrm{Tr}\left(\boldsymbol{V}\,\boldsymbol{W}^{-2}\boldsymbol{V}^T\right) = \sum_{k=1}^{n}\frac{\sigma^2}{w_k^2}\sum_{i=1}^{n}V_{ik}^2\end{aligned} \tag{A23}$$

where $\sigma_{x\,j}$ is the standard deviation of the estimate of reduced parameter x_j and $\sigma_{\beta\,j}$ the corresponding one for β_j. Since the right singular vectors have a unit norm ($\|\boldsymbol{V}_k\|^2 = \sum_{i=1}^{n} V_{ik}^2 = 1$), this last equation becomes:

$$\mathrm{Tr}\left(\mathrm{cov}(\hat{\boldsymbol{x}})\right) = \sum_{j=1}^{n}\left(\sigma_{\beta\,j} / \beta_j^{nom}\right)^2 = \sigma^2 \sum_{k=1}^{n} \frac{1}{w_k^2} \tag{A24}$$

In order to get a good estimation (in percent) of all the parameters of the model, the quadratic mean of the relative standard deviations of their estimates m_q should be smaller than a given level $m_{q\max}$ (NB: subscript q corresponds here to the quadratic mean of the normalized standard deviations):

$$m_q = \left(\frac{1}{n}\sum_{j=1}^{n}\left(\sigma_{\beta\,j} / \beta_j^{nom}\right)^2\right)^{1/2} = \sigma\left(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{w_k^2}\right)^{1/2} \le m_{q\max} \tag{A25}$$

One of the objectives of the "inverter" (the person in charge of the inversion) is to get a relative error m_q, expressed in term of quadratic mean, lower than an upper threshold $m_{q\max}$ equal to a few percent. This means that as soon as the number n of parameters that have to be estimated becomes large, the singular values w_k of the corresponding reduced sensitivity matrix decrease, which increases the error. This increase of the error is proportional to the standard deviation of the noise. This standard deviation has the same unit as the output of the signal and the same is true for the singular values which do not depend on the structure of the model (function $\boldsymbol{\eta}$) only, but also on the intensity of the stimulation (in a problem where the output is related to a field: temperature, concentration, …) and on the choice of the "times" of observation $\boldsymbol{t}$.

Both a lower and an upper level can also be constructed for the criterion of global relative error m_q defined in (A25), using the smaller singular value w_n:

$$\boxed{\frac{1}{\sqrt{n}}\frac{\sigma}{w_n} \le m_q = \left(\frac{1}{n}\sum_{j=1}^{n}\left(\sigma_{\beta\,j} / \beta_j^{nom}\right)^2\right)^{1/2} \le \frac{\sigma}{w_n}} \tag{A26}$$

This clearly shows that a too large value for the ratio σ / w_n, between the standard deviation of the measurement noise and the smaller singular value of the reduced sensitivity matrix $\boldsymbol{S}^*(\boldsymbol{\beta}^{nom})$, can make the estimation of the whole set of parameters « explode ». In that case, one of the β_j parameters (the parameter "supposed to be known", β_{sk}) has to be removed from the original set of parameters to be estimated. This will lead to a new parameter vector $\boldsymbol{\beta}'$ to be estimated, of smaller dimensions (n-1, 1), with a better (smaller) associated m_q criterion (lower average dispersion) but with the apparition of a bias on its n-1 estimates,

because of the biased value of the removed parameter β_{sk} that will be fixed to its nominal value that is different from its exact value (see Lecture 3).

Appendix 5 – Residual analysis for an unbiased model using the SVD approach

If the model used for estimation is unbiased, the residual vector, at convergence, is defined by:

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}) \equiv \boldsymbol{y} - \boldsymbol{y}_{mo}(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact}) + \boldsymbol{\varepsilon} - \boldsymbol{y}_{mo}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\varepsilon} - \boldsymbol{S}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{exact}) \text{ with } \boldsymbol{S} = \boldsymbol{S}(\hat{\boldsymbol{\beta}}) \quad \text{(A27)}$$

The last approximation in equation (A27) is based on a first order development of the model with respect to parameter $\boldsymbol{\beta}$, assuming that $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{exact}$ are close. So,

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\varepsilon} - \boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T(\boldsymbol{y} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact})) = \boldsymbol{\varepsilon} - \boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T(\boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact}) + \boldsymbol{\varepsilon} - \boldsymbol{y}_{mo}(\boldsymbol{\beta}^{exact})) \quad \text{(A28)}$$

The second term in equation (A28) is also a first order development that stems from the Gauss-Newton algorithm (12) used for minimizing $J_{OLS}(\boldsymbol{\beta})$ defined in (9) in an iterative way.

After simplification, equation (A28) can be rewritten using the scaled sensitivity matrix $\boldsymbol{S}^*$:

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{I}_m - \boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T)\,\boldsymbol{\varepsilon} = (\boldsymbol{I}_m - \boldsymbol{S}^*(\boldsymbol{S}^{*T}\boldsymbol{S}^*)^{-1}\boldsymbol{S}^{*T})\,\boldsymbol{\varepsilon} \quad \text{with} \quad \boldsymbol{S}^* = \boldsymbol{S}(\hat{\boldsymbol{\beta}})\,\mathrm{diag}(\hat{\boldsymbol{\beta}}) \quad \text{(A29)}$$

The lean SVD form (32b) (in the main body of this paper) of the scaled sensitivity matrix (see also Appendix 1) can be used then:

$$\boldsymbol{S}^*(\hat{\boldsymbol{\beta}}) = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V}^T \quad \text{(A30)}$$

This yield, using the orthogonality property of the right singular matrix $\boldsymbol{V}$:

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^T)\,\boldsymbol{\varepsilon} \quad \text{(A31)}$$

So, under the IID noise assumption, for an unbiased model, one can show that the expectation of the residual vector is equal to-zero:

$$\mathrm{E}(\boldsymbol{r}(\hat{\boldsymbol{\beta}})) \approx (\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^T)\,\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0} \quad \text{(A32)}$$

This means that if the model used for describing the experiment is appropriate, the residuals curve is centred on the $y = 0$ axis.

In order to get "unsigned" residuals, the variance-covariance matrix of the residuals should be diagonal. If the model is unbiased, this matrix is:

$$\mathrm{cov}(\boldsymbol{r}(\hat{\boldsymbol{\beta}})) \approx (\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^T)^T\,\mathrm{cov}(\boldsymbol{\varepsilon})\,(\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^T) = \sigma^2(\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^T) = \sigma^2\,\boldsymbol{U}_{comp}\boldsymbol{U}_{comp}^t \quad \text{(A33)}$$

Here $\boldsymbol{U}_{comp}$ is the complementary left singular vectors matrix composed of the (m – n) left singular vectors, that appear in the full SVD decomposition of $\boldsymbol{S}^*(\hat{\boldsymbol{\beta}})$ given by equation (A1b) in *Appendix 1*:

$$\boldsymbol{S}^* = \boldsymbol{U}_0\ \boldsymbol{W}_0\ \boldsymbol{V} \quad \text{with} \quad \boldsymbol{U}_0 = \left[\boldsymbol{U}\ \ \boldsymbol{U}_{comp}\right] \text{ where } \boldsymbol{U}_0^T\boldsymbol{U}_0 = \boldsymbol{I}_m \text{ and } \boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{0}_{(m-n)\ x\ n} \end{bmatrix} \qquad \text{(A34)}$$

In case of a square non-linear least square problems, there are as many measurements as parameters to be estimated ($m = n$) and $\boldsymbol{I}_m = \boldsymbol{U}\boldsymbol{U}^T$. So, in this case, the residuals (A27) are deterministic and equal to zero ($\boldsymbol{U}_{comp}$ is an 'empty' matrix with 0 column in that degenerated case). As soon as the number m of measurements gets higher than the number n of parameters, matrix $\boldsymbol{U}_{comp}\boldsymbol{U}_{comp}^t$ becomes non-diagonal, especially if the difference (m – n) is small and the residuals are correlated. However, when this difference increases, that is when the number of measurements is a lot higher than the number of parameters, the ratio n/m goes to zero and $\boldsymbol{U}_{comp}$ becomes very close to $\boldsymbol{U}_0$, which means that

$$\text{cov}(\boldsymbol{r}(\hat{\boldsymbol{\beta}})) \approx \sigma^2\ \boldsymbol{U}_{comp}\boldsymbol{U}_{comp}^t \xrightarrow[\text{as } n/m \to 0]{} \sigma^2\ \boldsymbol{U}_0\boldsymbol{U}_0^T = \sigma^2\boldsymbol{I}_m \qquad \text{(A35)}$$

This means that, strictly speaking, the residuals are correlated, even for an unbiased model but, in practice, adding more many measurement times to a given estimation interval tends to make them nearly uncorrelated. This is especially true for thermal characterization of materials or system, where the number of parameters is low (2, 3, 4, ...) and the time sampling rate high enough with respect of the length of measurement (several hundred measurements at least for modern data acquisition systems) where the asymptotic level given by (A35) is reached.

Lecture 6. Inverse problems and regularized solutions

J.C. Batsale[1], O. Fudym [2]
Paper of METTI6 revised by C. Le Niliot[3]
[1] I2M Lab., Departement TREFLE, Université Bordeaux 1, CNRS, ENSAM, Bordeaux-INP, France
e-mail: jean-christophe.batsale@u-bordeaux.fr
[2] Université de Toulouse ; Mines Albi; centre RAPSOODEE, Albi, France
e-mail: fudym@mines-albi.fr
[3] IUSTI, Aix Marseille Univ. UMR 7343 CNRS, Marseille, France
e-mail : Christophe.LeNiliot@univ-amu.fr

Abstract. The methods for solving inverse problems must propose some consistent solutions despite their ill-posed character. Regularization is one of the major techniques for stabilizing the solution. We present in this lecture some generic examples as well as the main concepts within the linear estimation frame for the OLS (Ordinary Least Squares) estimator already studied in Lectures 1 and 3. The Singular Value Decomposition of the sensitivity matrix is used in order to analyse the solution. For such finite dimensional problems, the ill-posed behavior results in a bad-conditioned matrix computation.

1. Introduction

The reader could see in Lecture 1, "Getting started with problematic inversions with three basic examples", some examples of generic inverse problems, which gave rise to envision the main characteristics that make their solution difficult. In Lecture 3, "Basics for linear inversion, the white box case", the concepts and resolution of linear parameter estimation problems were presented, when using a direct model that computes the output from the knowledge of the input and some inner parameters used in the direct model.
The parameters to be recovered may be as well the passive structural parameters of the model (model identification), the parameters relative to the input variables, initial state, boundary conditions, some thermophysical properties, calibration, etc… For any of these considered cases, the output of the model can be properly computed if all the required information is available.

In a famous book, Hadamard (Hadamard 1923) introduced in 1923 the notion of well-posed problem. This is a problem whose solution:
- exists;
- is unique;
- depends continuously on the data.

Of course, these notions must be specified by the choice of space (and topologies) in which the data and the solution evolve. Problems that are not well-posed in the sense of Hadamard are said to be *ill-posed problems*. Note that the simple inversion of a well-posed problem may be either a well-posed or an ill-posed problem.

The example of 1D steady heat conduction in a wall discussed in Lecture 1, shows how the interpolation problem (that is the computation of $T(x)$ between the sensor location and the well-known boundary condition) is a well-posed problem, while the extrapolation problem (computation of $T(x)$ between the unknown boundary condition to be retrieved and the sensor location) is an ill-posed problem, since the estimation error may increase drastically.

The example of searching the slope of a line with two or more data points, such as discussed in Lecture 3, may be either a well-posed or an ill-posed problem:

- a unique and stable solution exists if all the data points fit on the same line (no noise in the data), and the time zero has not been chosen for some noisy data point. In that very specific case, the problem of finding the slope is well-posed.
- If, due to the noise in the measurement points, the data do not fit on the same line, a solution does not exist and the corresponding inverse problem of finding the slope is ill-posed.
- If the values of time for taking the measurements are not properly chosen (mostly close to zero), the solution is unstable, since the errors in the measurement may increase drastically – see the absolute and relative amplification coefficients such as defined in Lecture 1, and the corresponding inverse problem is ill-conditioned and may be considered as ill-posed.

The parameter estimation problem that consists in finding the vector of parameters by matching the measurements to the model outputs is most often an ill-posed problem, since it is generally over-determined (because the number of measurements m is greater than the number of parameters n), and has no solution because $\mathbf{y} \notin \mathrm{Im}(\mathbf{S})$. When the system is under-determined ($m < n$), it is also ill-posed because there is an infinite number of solutions. Moreover, when $m = n$, the problem may be well-posed if it were stable, but may also be unstable due to the effect of noise in the data.

In the present lecture, we will consider discrete inverse problems, where the number of parameters to be estimated is finite. When the quantities to be estimated are functions instead of discrete values, the corresponding problem is a continuous inverse problem which may be fully ill-posed. However, in many cases, the searched functions can be parametrized and conveniently approximated by a discrete inverse problem. It was typically the case for the 1D transient inverse heat conduction example in section 4 of Lecture 1, where the wall heat flux was to be estimated as a function of time. The heat flux at each time $\boldsymbol{t_i}$ is represented by a stepwise function $\boldsymbol{q_i}$.

The main challenge for such discrete function estimation problem is that the number of unknown is almost the same as the number of measurements, and the least squares approach is quite close to an exact matching procedure where only one observation is available for one estimated value. In this case the solution is highly sensitive to any ill-conditioned behaviour of the sensitivity matrix.

2. Some examples of typical ill-posed problems

We give hereafter some typical examples of ill-posed problems, such as derivation and deconvolution of experimental data. These examples are typical of the case of a parameterized

function estimation problem. Instead of having a low number of parameters to be estimated with a high number of measurements, as for the example in Lecture 3 for estimating the slope and intercept of a straight line, the number of parameters to be estimated is herein very large and is quite of the same order as the number of observable data $\boldsymbol{y}$, which makes the problem highly sensitive to noise. Unfortunately, in this case the inversion is often also amplifying the measurement noise.

2.1 Derivation of a signal

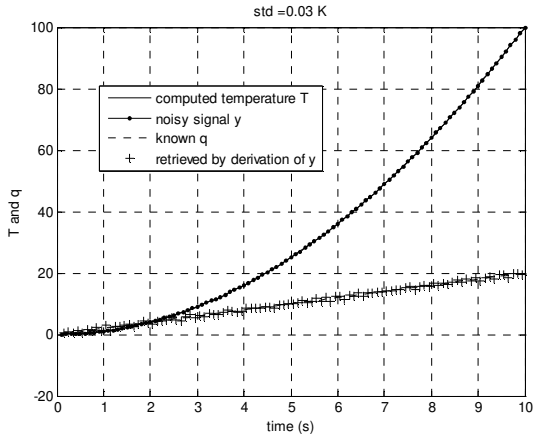
The derivation of a signal is often required for data processing. It is the case of time dependent functions, for instance, when deriving the time evolution of the mass of a product during drying or deducing the velocity of a body from the measurement of its position. A usual case in heat transfer is the problem of estimating the heat flux $q(t)$ exchanged by a body with uniform temperature $T(t)$ and volumetric heat capacity (ρC), with the lumped body approximation for a small thickness e . The heat balance can be written as

$$(\rho Ce)\frac{dT}{dt}=q(t) \qquad \text{with the initial condition } t=0 \quad T=0 \tag{6.1}$$

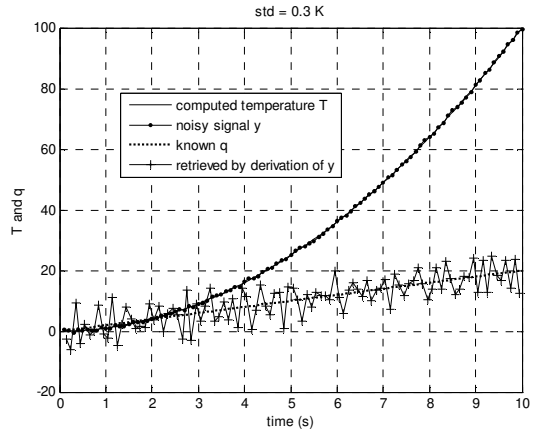
An inversion procedure is sought, for recovering an estimation of $q(t)$ from the measured temperature values $y(t_i)$, for different levels of the measurement noise, based on the following steps:

a. Choose some heat flux function, such as $q(t)=2t$ (arbitraly chosen here)
b. Compute the corresponding analytical solution $T(t)=t^2/(\rho Ce)$.
c. Add some random error, in order to simulate some experimental data, such as $y(t)=T(t)+\varepsilon(t)$
d. Retrieve the estimation by discrete derivation of the signal $\hat{q}(t)=(\rho Ce)\frac{\Delta y}{\Delta t}\approx\ (\rho Ce)\frac{dT}{dt}$
e. Repeat for different values of the Signal-to-Noise Ratio (characterized by different levels of standard deviation (std))

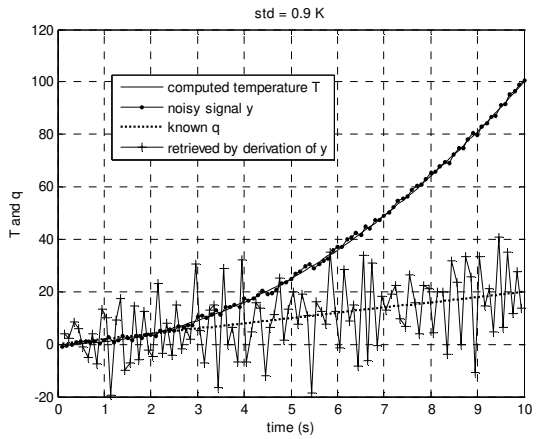
The results are depicted in Figure 1, assuming that $(\rho Ce)=1$. When the standard deviation of the error is low, the heat flux is conveniently retrieved (Fig. 1a). For case (b), the noise on the signal y remains very low, in the sense it is still almost not visible in the corresponding curve. However, the heat flux is poorly computed. Increasing the level of noise, such as in Fig. 1c, where the std is *0.9 K*, results in a drastically poor computation of the heat flux. Thus, the numerical derivation of an experimental signal in order to retrieve its integral is an ill-posed problem, due to its unstable nature. The numerical derivation consists in computing the difference of successive measurements, divided by the time step. In this case the ill-posed character of the problem could be more dramatic as the time step decreases. As a result, the fluctuations in the identified function could be as important as we would like if Δt tends to 0. It is an illustration of the ill-posed character of the inverse problems: a bounded error can be amplified to infinity and the third condition of Hadamard is not satisfied!



(a)



(b)



(c)

Figure 1 – Derivation of an experimental signal (a) std = 0.03 K (b) std = 0.3 K (c) std = 0.9 K

2.2 Deconvolution of a signal

The deconvolution of a signal is also an operation often required when processing experimental data, for instance when searching the transfer function of a system or sensor, in image processing, optics, geophysics, etc... We give again the heat transfer example of some heat capacity exchanging with convective heat losses with the surrounding medium, such as

$$(\rho Ce)\frac{dT}{dt} = q(t) - hT \quad \text{with the initial condition } t = 0 \quad T = 0 \tag{6.2}$$

We assume here that $(\rho Ce) = 1, h = 1$ and that the area of the boundary surface of the body is 1.

Solving this equation by using the Laplace transformation of the temperature and heat flux with an analytical return to the time domain yields the solution in the form of the following convolution product:

$$T(t) = \int_0^t q(t-\tau)\, exp(-h\tau)\, d\tau = \int_0^t q(\tau)\, exp(-h\,(t-\tau))\, d\tau \tag{6.3}$$

The same approach as in previous example is proposed herein, such as

a.Choose some heat flux function, $q(t)$, called the input. Here the input function is chosen as:

$q(t) = q_0\, exp\left(-\left(\frac{t-t_0}{\tau}\right)^2\right)$ with $q_0 = 10$; $t_0 = 0.5$ and $\tau^2 = 0.05$

b. Compute the corresponding analytical solution, that is the output $T(t)$, of the convolution product above.

c. Add some random error, such as $y(t) = T(t) + \varepsilon(t)$

d. Retrieve the heat flux by inverting the convolution product of this signal by the negative exponential above, that is the corresponding impulse response in this example (a numerical deconvolution). The impulse response can be noted as: $Imp(t) = exp(-h\,t)$.

e. Repeat for different values of the Signal-to-Noise Ratio (different levels of std: σ_ε)

The discrete approximation of expression (6.3) can be considered as a linear matricial expression between an input vector : $\boldsymbol{q}=[q_0\ q_1 \ldots q_{n-1}]$ and an output vector $\boldsymbol{T}=[T_1\ T_2 \ldots T_n]$, such as

$$\begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix} = \Delta t \begin{bmatrix} Imp_1 & 0 & \vdots & 0 \\ Imp_2 & Imp_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ Imp_n & \ldots & Imp_2 & Imp_1 \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{n-1} \end{bmatrix} \quad \text{or} \quad T_i = \Delta t \sum_{j=0}^{n-1} Imp_{i-j} * q_j \tag{6.3 bis}$$

With $Imp_k = \exp(-ht_k)$ and $t_k = k * \Delta t$ $k = 1\ to\ n$

The sensitivity matrix which here depends on only one vector: $\boldsymbol{Imp}=[Imp_1\ Imp_2 \ldots Imp_n]$ is called a lower triangular Toeplitz matrix, such as $\boldsymbol{S} = \boldsymbol{Toeplitz}(\boldsymbol{Imp})\Delta t$. Such a matrix is diagonal-constant.

The results from the Matlab script given in Appendix 1 are depicted in Figure 2. For a low standard deviation of the error (*std = 0.01 K*), the heat flux is conveniently retrieved by the deconvolution operation (simple inversion of the Toeplitz matrix). When increasing the noise level (*std = 0.1 K*), the drastic amplification of the errors in the deconvolution operation makes the result absolutely inaccurate. The increase of the noise level, that can be observed in the temperature plots, makes the solution (the input) inaccurate or even unavailable. This example shows that deconvolution of an experimental signal may be an ill-posed problem, depending on the functional form of the impulse response and on the noise level.

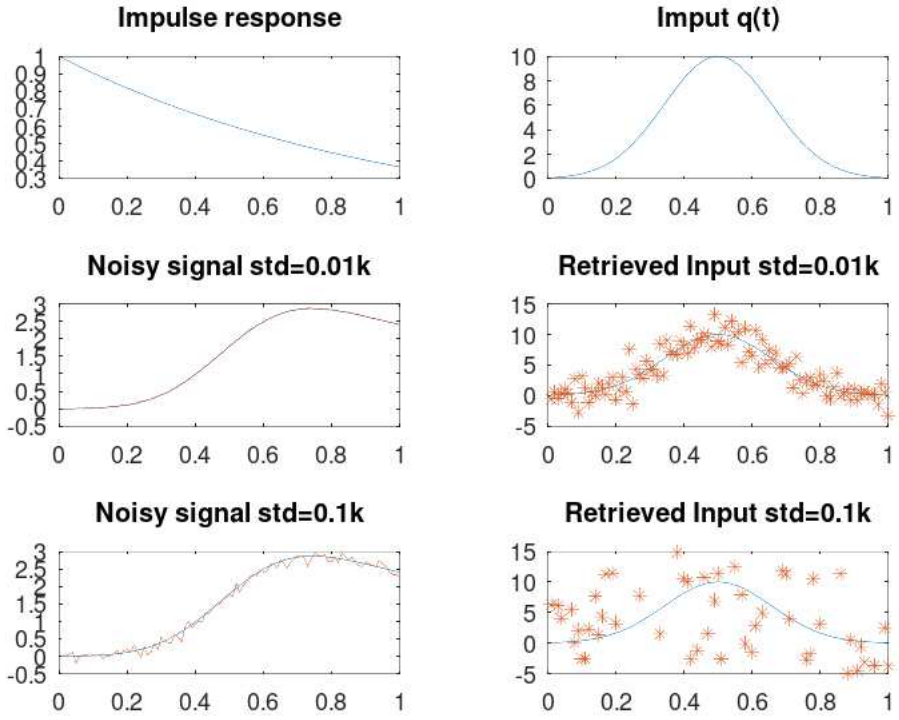


Figure 2 – Effect of the noise level on the deconvolution of a signal
(continuous blue line: exact signal, red line or dotted: noisy or retrieved signal)

3. Structure of a linear transformation and stability

3.1 Singular Value Decomposition of the sensitivity matrix

It was already shown in Lecture 3 of this school that the existence, uniqueness and stability of the solution of a discrete linear parameter estimation problem depends on the characteristics and structure of the rectangular sensitivity matrix $\boldsymbol{S}$. Moreover, when the overdetermined problem $\boldsymbol{y}=\boldsymbol{S}\boldsymbol{x}$ is considered as a least square problem given by the normal equations, it appears that the structure of the square information matrix $\boldsymbol{S}^t\boldsymbol{S}$ has a major effect on the

propagation of the errors between the observed data and the output of the model. The anatomy of such a linear transformation is very clearly discussed in the text of S. Tan & C. Fox (Tan 2006).
One approach of interest in order to analyze this problem is to consider the Singular Value Decomposition of $\boldsymbol{S}$ (SVD). We assume herein that $m > n$ (overdetermined system, there is more data than parameters) and that $\boldsymbol{S}$ has only real coefficients. All the details of SVD analysis can be found in (Hansen 1998) and the corresponding routines are available in any of the linear algebra librairies (LAPACK, Num. Recipes, Matlab®, …).
The SVD of the matrix $\boldsymbol{S}$ is then written as

$$\begin{bmatrix} \\ \boldsymbol{S} \\ \\ \end{bmatrix} = \boldsymbol{U}\,\boldsymbol{W}\,\boldsymbol{V^t} = \begin{bmatrix} \\ \boldsymbol{U} \\ \\ \end{bmatrix} \begin{bmatrix} W_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & W_n \end{bmatrix} \begin{bmatrix} \\ \boldsymbol{V^t} \\ \\ \end{bmatrix} \tag{6.4}$$

where

- $\boldsymbol{U}$ is an orthogonal matrix of dimensions *(m,n)* : its column vectors (the *left* singular vectors of $\boldsymbol{S}$) have a unit norm and are orthogonal by pairs : $\boldsymbol{U^t U = I_n}$, where $\boldsymbol{I_n}$ is the identity matrix of dimension n. Its columns are composed of the first *n* eigenvectors U_k, ordered according to decreasing values of the eigenvalues of matrix $\boldsymbol{S S^t}$. Let us note that, in the general case, $\boldsymbol{U U^t \neq I}_m$.

- $\boldsymbol{V}$ is a square orthogonal matrix of dimensions *(n,n)* , : $\boldsymbol{V V^t = V^t V = I_n}$. Its column vectors (the *right* singular vectors of $\boldsymbol{S}$), are the n eigenvectors $\boldsymbol{V_k}$, ordered according to decreasing eigenvalues, of matrix $\boldsymbol{S^t S}$;

- $\boldsymbol{W}$ is a square diagonal matrix of dimensions *(n,n)*, that contains the n so-called singular values of matrix $\boldsymbol{S}$, ordered according to decreasing values : $\boldsymbol{W_1 \geq W_2 \geq \cdots W_n}$ The singular values of matrix $\boldsymbol{S}$ are defined as the square roots of the eigenvalues of matrix $\boldsymbol{S^t S}$.

In Lecture 3, the Singular value Decomposition of the reduced sensitivity matrix, through the analysis of its singular values, was used to demonstrate that its condition number is a criterion that can be used to measure the degree of ill-posedness of the OLS estimator, regardless of the noise level.

As previously seen in Lecture 3, the Ordinary Least Squares solution is obtained by minimizing the distance between the output of the direct model $\boldsymbol{Sx}$ and the data $\boldsymbol{y}$, which is done by the orthogonal projection of the data on the space spanned by the column vectors of $\boldsymbol{S}$. This is equivalent to minimizing the objective function

$$J_{OLS}(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{Sx}\|^2 = (\boldsymbol{y} - \boldsymbol{Sx})^{\boldsymbol{t}}\,(\boldsymbol{y} - \boldsymbol{Sx}) \tag{6.5}$$

The minimization of $J_{OLS}(\boldsymbol{x})$ yields the OLS estimator, computed with Eq. (3.24) in Lecture 3. Applying the Singular Value Decomposition to the sensitivity matrix yields

$$\hat{x}_{OLS} = (\boldsymbol{S^t S})^{-1}\,\boldsymbol{S^t y} = \boldsymbol{V}\ \boldsymbol{W^{-1}}\,\boldsymbol{U^t y} \tag{6.6}$$

In this case, if the standard statistical assumptions hold (see Lecture 3), the covariance matrix of the OLS estimator can be written as

$$cov(\boldsymbol{x}) = {\sigma_\varepsilon}^2 \, \boldsymbol{V} \; \boldsymbol{W^{-2}} \, \boldsymbol{V^t} \tag{6.7}$$

Eqs. (6.6) and (6.7) are valid if the sensitivity matrix $\boldsymbol{S}$ is of full rank, which means that its smaller singular value w_n is strictly positive. The condition number is then defined as

$$cond(\boldsymbol{S}) = \frac{w_1}{w_n} \tag{6.8}$$

3.2 Spectral analysis of the OLS estimator

Applying SVD to the normal equations (see Eq. (3.23) in Lecture 3) in order to find the OLS estimator in the diagonal basis yields

$$(\boldsymbol{S^t \, S}) \; \hat{\boldsymbol{x}}_{\boldsymbol{OLS}} = \; \boldsymbol{S^t \, y} \;\; \Rightarrow \;\; \boldsymbol{V \;\; W \, U^t \, U \, W \, V^t} \, \hat{\boldsymbol{x}}_{\boldsymbol{OLS}} = \; \boldsymbol{V \;\; W \, U^t \, y} \tag{6.9}$$

where the estimation problem can be reconsidered now with the new parameter vector $\boldsymbol{b} = \; \boldsymbol{V^t x}$ and a new observable vector : $\boldsymbol{z} = \; \boldsymbol{U^t y}$, such as

$$\boldsymbol{W} \; \hat{\boldsymbol{b}}_{\boldsymbol{OLS}} = \; \boldsymbol{z} \tag{6.10}$$

The unicity of the solution is confirmed here when the sensitivity matrix $\boldsymbol{S}$ is of full rank, i.e. $r = n$, which is possible only if $m \geq n$ (more data than parameters). When $r < n$, the matrix has not full rank, and the number of parameters to be estimated must be reduced, or some parameters must be determined in an arbitrary form.

The linear transformation of the data $\boldsymbol{y}$ also yields a new covariance matrix associated to the observable measurement noise. Hopefully, we can note that this operation does not affect the covariance of the error of the transformed signal $\boldsymbol{z}$ (here for the standard assumptions):

$$cov(\boldsymbol{z}) = \boldsymbol{U^t} cov(\boldsymbol{y}) \, \boldsymbol{U} = \; {\sigma_\varepsilon}^2 \, \boldsymbol{U^t \, U} = \; {\sigma_\varepsilon}^2 \boldsymbol{I} \tag{6.11}$$

Hence the covariance matrix of the error of $\hat{\boldsymbol{b}}_{\boldsymbol{OLS}}$ is computed by

$$cov\left(\hat{\boldsymbol{b}}_{\boldsymbol{OLS}}\right) = \; {\sigma_\varepsilon}^2 \, \boldsymbol{W^{-2}} \qquad \text{or} \;\; cov\left(\hat{\boldsymbol{b}}_{\boldsymbol{OLS}}\right) = \begin{bmatrix} \frac{{\sigma_\varepsilon}^2}{{w_1}^2} & . & \boldsymbol{0} \\ . & . & . \\ \boldsymbol{0} & . & \frac{{\sigma_\varepsilon}^2}{{w_n}^2} \end{bmatrix} \tag{6.12}$$

The above equation shows that an effect of noise amplification appears due to the fact that the eigenvalues have a wide range of orders of magnitude. It is of particular interest to note in Eq. (6.12) that the covariance matrix of the estimator in the diagonal basis is linking the square of the singular values to the variance of noise, that is to the level of uncertainty in the measurement errors.

A small perturbation applied to a single component k of $\boldsymbol{z}$, such as

$$\boldsymbol{\delta z} = \delta z_k \boldsymbol{U_k} \tag{6.13}$$

yields the following variation to the OLS estimator

$$\boldsymbol{\delta \widehat{b}}_{\boldsymbol{OLS}} = \frac{\delta z_k}{w_k} \boldsymbol{V_k} \tag{6.14}$$

which implies a relative variation corresponding to

$$\frac{\|\boldsymbol{\delta \widehat{b}}_{\boldsymbol{OLS}}\|}{\|\boldsymbol{\delta z}\|} = \frac{1}{w_k} \tag{6.15}$$

Thus the singular values indicate how the same perturbation yields different effects on the components of the estimator. Moreover, this relative variation may increase drastically when the singular values are close to zero. The relative variation between two components of respective index k and h is given by the ratio $\frac{w_k}{w_h}$. Hence the maximum relative variation factor is obtained between the first and the last component, such as $\frac{w_1}{w_n} = cond(\boldsymbol{S})$, which is the condition number of the sensitivity matrix, as seen in Eq. (6.8). If is not too large, the problem is said to be well-conditioned and the solution is stable with respect to small variations of the data. Otherwise the problem is said to be ill-conditioned. It is clear that the separation between well-conditioned and ill-conditioned problems is not very sharp and that the concept of well-conditioned problem is more vague than the concept of well-posed problem.

3.3 Example of a simple ill-conditioned matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.01 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{the inversion yields } \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Let's give a perturbation of 1% on the second data point, such as

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.01 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.01 \end{bmatrix} \qquad \text{the inversion yields } \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Hence the perturbation of the data makes the solution of the matrix inversion, that is the solution of the square linear system of 2 equations with two unknowns, surprisingly as far as possible from the original solution. The solution is quite unstable.

The eigenvalues are (2.005, 0.005), and the condition number is 402 >> 1.

4. Regularization

In the previous section it was shown how the ill-posed estimation problem is turned into an ill-conditioned problem by the least squares approach. Equations. (6.6) and (6.7) show that the unstable behavior of the pseudo-inverse of the sensitivity matrix can be straightly addressed by the means of the singular values diagonal matrix $\boldsymbol{W}$. Regularization is a process for searching some acceptable solution, by reducing the effect of measurement errors on the estimate. Several approaches may be used for this purpose. The main idea is to reduce the effect of the "small" singular values on the obtained solution, while trying to avoid that this

"smooth" solution to be too different from the true-but-untractable solution. The two main approaches are

(i) *Apply some prior information as a constraint*
The most usual methods are truncation of the diagonal basis and the parametrization of the solution (reduction of the number of parameters)

(ii) *Apply some penalization to the objective function*
Some weighted prior information is included in the objective function

4.1 Truncated SVD (TSVD)

TSVD of order α is obtained through replacing in Eq. (6.6) the inverse of the matrix $\boldsymbol{W}$ by the truncated inverse $\boldsymbol{W_\alpha}^{-1}$ where the smaller singular values are removed (their inverse put to zero), such as

$$\boldsymbol{W_\alpha}^{-1} = \begin{bmatrix} 1/w_1 & & & 0 & & \\ & 1/w_2 & & & & \\ & & \ddots & & & \\ & & & 1/w_\alpha & & \\ & 0 & & & 0 & \\ & & & & & \ddots \\ & & & & & 0 \end{bmatrix} \tag{6.16}$$

The regularized TSVD estimator is:

$$\widehat{\boldsymbol{x}}_{\boldsymbol{\alpha}}^{\boldsymbol{TSVD}} = \boldsymbol{V}\ \boldsymbol{W_\alpha}^{-1}\ \boldsymbol{U^t}\ \boldsymbol{y} \tag{6.17}$$

We note that only $\boldsymbol{W_\alpha}^{-1}$ can be computed, and not matrix $\boldsymbol{W_\alpha}$ since an infinite value has been assigned to the $n-\alpha$ smallest singular values $W_{\alpha+1}, W_{\alpha+2}, \ldots, W_{n-1}, W_n$.

Equation (6.17) may be written using the left and right singular column vectors $\boldsymbol{U_k}$ and $\boldsymbol{V_k}$ defined in section 1, such as

$$\widehat{\boldsymbol{x}}_{\boldsymbol{\alpha}}^{\boldsymbol{TSVD}} = \sum_{k=1}^{\alpha} \frac{1}{w_k}\ \left(\boldsymbol{U_k}^{\boldsymbol{t}}\boldsymbol{y}\right)\boldsymbol{V_k} \tag{6.18}$$

The choice of the truncation order α can be considered by taking into account the observable measurement noise such as:

$$W_\alpha \geq \sigma_\varepsilon > W_{\alpha+1} \tag{6.19}$$

Unfortunately, it is not always easy to determine the standard deviation of the measurement noise for example in the case of non-stationary signal. The choice of the truncation order without the knowledge of σ_ε will be discussed in section 4.4.

4.2 Tikhonov regularization of zero order

The important idea in introducing some regularization by penalizing the objective function is the will to include some prior knowledge relative to the parameters to be retrieved. For instance, the parameter should not be very far from a reference value, or the time history of the function to be estimated should be smooth… A widespread regularization method by penalization of the OLS objective function is Tikhonov regularization.

We present herein the Tikhonov regularization of order zero, which yields the minimization of the following objective function:

$$J_\mu(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}\|^2 + \mu\left\|\boldsymbol{x} - \boldsymbol{x}_{prior}\right\|^2 = (\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x})^{\boldsymbol{t}}(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}) + \mu\left(\boldsymbol{x} - \boldsymbol{x}_{prior}\right)^{\boldsymbol{t}}\left(\boldsymbol{x} - \boldsymbol{x}_{prior}\right) \quad (6.20)$$

where the real positive number μ is the regularization parameter. The value μ =0 yields the OLS solution where no regularization applies. Increasing μ tends to force the solution to be close to the prior estimate $\boldsymbol{x}_{prior}$

Equation (6.20) is solved by:

$$\hat{\boldsymbol{x}}_{\boldsymbol{\mu}}^{Tik0} = (\boldsymbol{S}^{\boldsymbol{t}}\,\boldsymbol{S} + \mu\boldsymbol{I}_{\boldsymbol{n}})^{-1}\left(\boldsymbol{S}^{\boldsymbol{t}}\,\boldsymbol{y} + \mu\boldsymbol{x}_{prior}\right) \quad (6.21)$$

Applying SVD to the sensitivity matrix $\boldsymbol{S}$ and using $\boldsymbol{V}\boldsymbol{V}^{\boldsymbol{t}} = \boldsymbol{I}_{\boldsymbol{n}}$ yields:

$$\hat{\boldsymbol{x}}_{\boldsymbol{\mu}}^{Tik0} = \boldsymbol{V}\left(\boldsymbol{W}^{\boldsymbol{2}} + \mu\boldsymbol{I}_{\boldsymbol{n}}\right)^{-1}\left(\boldsymbol{W}\boldsymbol{U}^{\boldsymbol{t}}\,\boldsymbol{y} + \mu\boldsymbol{V}^{\boldsymbol{t}}\boldsymbol{x}_{prior}\right) \quad (6.22)$$

Equation (6.22) clearly shows that the regularization parameter will cancel the noise amplification effect of the smallest singular values in the diagonal matrix $\left(\boldsymbol{W}^{\boldsymbol{2}} + \mu\boldsymbol{I}_{\boldsymbol{n}}\right)$ to be inverted. Nevertheless, the cost of this stabilization is also obvious, since the non-zero regularization parameter value yields that the information of the experimental data in $\boldsymbol{y}$ is biased by the prior information ($\boldsymbol{x}_{prior}$). Hence let's point out that the regularized solution aims to balance accuracy and stability requirements.

4.3 Example: Regularization for deconvolution

The experimental derivation and deconvolution example given in section 2 can be solved as a linear estimation problem. The function estimation problems are highly sensitive to noise, since the number of unknown matches the number of function components to be retrieved (exact matching: the sensitivity matrix is a square matrix).

Expression (6.17) (TSVD method) and (6.21) (Thikonov method) are used for the regularization of the output of the model given by equations (6.2) and (6.3), for $h = \rho C e$ =1 and with the same input as the one presented in figure 2.

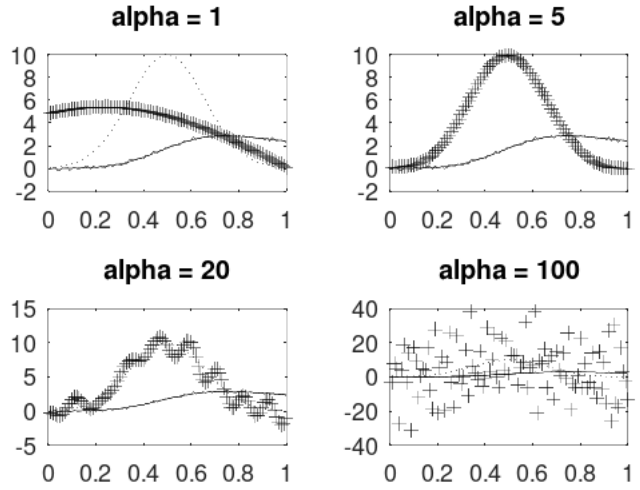


Figure 3 – Deconvolution and inversion with TSVD regularization
— temperature (exact and noisy); - - exact input ; + estimated input.

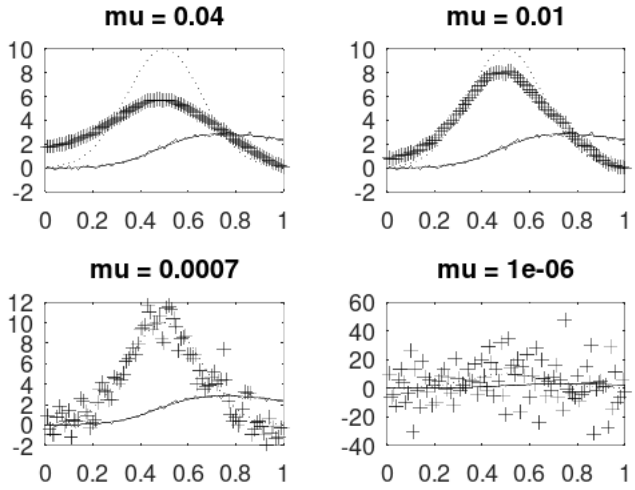


Figure 4 – Deconvolution and inversion with Tikhonov regularization
— temperature (exact and noisy); - - exact input ; + estimated input.

Figure 3 and 4 show how increasing the value of the regularization parameter has a positive effect regarding the stabilization of the heat flux time history to be retrieved, while this effect is counter balanced by the apparition of a bias with the original solution. It is of great interest to point out that the correct possible values of the regularization parameter are related to the signal to noise ratio. In the case of the TSVD method, the truncation parameter α is near 5 and the corresponding singular values are : $w_3 = 0.124$ $w_4 = 0.898$ and $w_5 = 0.706$. In the case of the Thikonov method, the regularisation parameter μ is quite close to the variance of the measurement error (here $\sigma_\varepsilon{}^2$=0.01K).

The Matlab codes related to the figures 3 and 4 are given in appendix 2 and 3.

4.4 The regularization parameter

The optimal choice of the value of the regularization parameter is a nontrivial problem for which numerous solutions have been proposed. Such problem is accentuated if the variance of the measurement noise is poorly known. The L-curve method (due to Hansen,1998) has become a popular method, which is implemented by the graphical analysis of a log-log plot (or ordinary plot) obtained by varying the value of the regularization parameter, as shown in figure 5. For each value of μ, the norm of the distance between the data and the model is reported on the horizontal axis, while the distance of $\boldsymbol{x}$ to $\boldsymbol{x_{prior}}$ is reported on the vertical axis. Very often the vector $\boldsymbol{x_{prior}}$ is set to zero initially. An iterative process can be further implemented. The L-curve selection criterion consists in locating the value which maximizes the curvature, that is the L-curve corner which separates the two regions: under-regularized on the left, over-regularized on the right

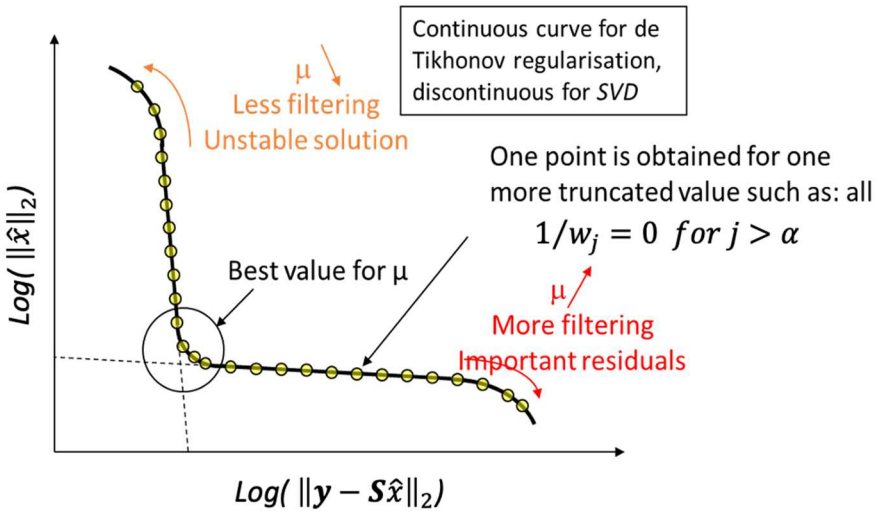


Figure 5 – L Curve, choice of the Tikhonov regularization parameter μ, comparison with truncated SVD solution

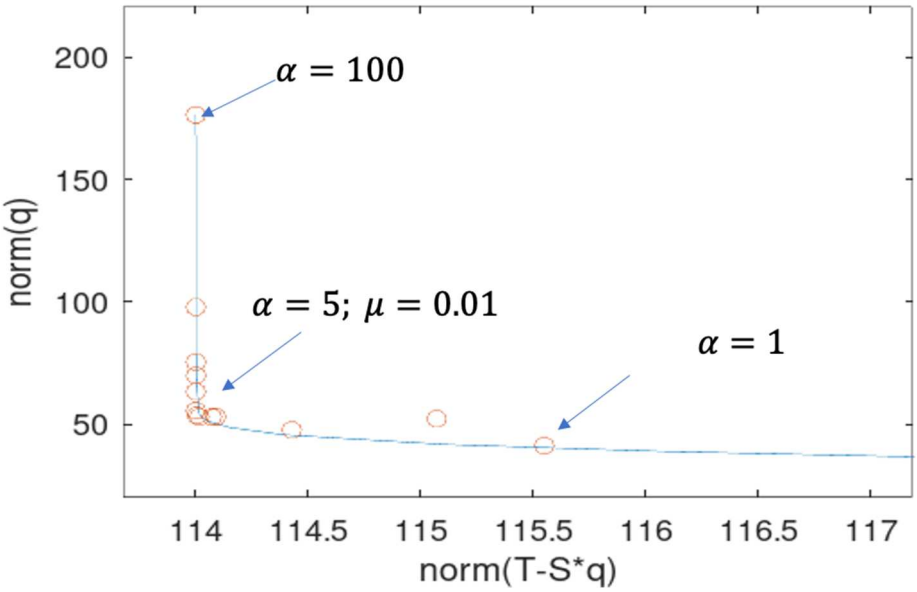


Figure 6 – L-curve obtained for the regularization of the deconvolution example

The L-curve is plotted for the previous regularized deconvolution example on figure 6. The red circles are the regularisation parameters obtained with the TSVD method (α from 1 to n). The continuous line is related to the parameter μ. The best regularization parameter μ is near the same value as the variance of the measurement noise. The best truncation number α is near 5 . It is difficult to distinguish the results with $\alpha = 4$ or $\alpha = 6$. All these points are very close to the point that maximizes the curvature of the L-curve.

The Matlab code related to figure 6 is given in appendix 4.

6. Conclusions

Regularization is an important step for solving ill-posed problems. When the inverse problem is of finite dimension, which is the case for discrete estimation problems, the existence of a solution is achieved by the least squares approach, and the problem is in fact ill-conditioned. For function estimation problems, the parametrization of the function to be retrieved tends to exact matching, where the number of experimental data is equal to the number of parameters. This case is generally highly sensitive to the measurement noise. Regularization stabilizes the solution by removing the effect of the smallest singular values which amplify the effect of these measurement errors. However the cost of regularization is a biased stabilized solution, hence the value of the regularization parameter (Tikhonov parameter or truncation level) must be carefully chosen.

References

Hadamard, J. 1923. Lectures on Cauchy's Problem in Linear Differential Equations, Yale University Press, New Haven, CT.

Aster R. C., Borchers B., Thurber C. H., *Parameter Estimation and Inverse Problems*, Elsevier Academic Press, 2005

Beck, J. V. and Arnold K. J., *Parameter estimation in engineering and science*, John Wiley & Sons, 1977

THERMAL MEASUREMENTS AND INVERSE TECHNIQUES, Edited by Helcio R. B. Orlande, Olivier Fudym, Denis Maillet, Renato M. Cotta, CRC Press, New York, ISBN : 978-1-4398-4555-4, 2011

Ozisik, M.N. and H.R.B. Orlande. 2000. Inverse Heat Transfer: Fundamentals and Applications, Taylor and Francis, New York.

Campbell, S. L. and C. D. Jr Meyer. 1991. Generalized Inverses of Linear Transformations. New York: Dover.

Hansen, P.C., 1998. Rank Deficient and Ill-Posed Problems: Numerical Aspects Of Linear Inversion, SIAM.

Jenkins, G.M. and D.G. Watts. 1998. Spectral analysis and its applications, Emerson-Adams Press.

Tan, S., C. Fox and G. Nicholls. 2006. Inverse Problems, Course Notes for Physics 707, University of Auckland, 2006. http://www.math.auckland.ac.nz/%7Ephy707/

Tikhonov, A. and V. Arsénine. 1976. Méthodes de résolution des problèmes mal-posés, Editions de Moscou.

Shenfelt, J.R., R. Luck, R.P. Taylor and J.T. Berry. 2002. Solution to inverse heat conduction problems employing SVD and model reduction. IJHMT 45:67-74.

Fudym, O., C. Carrère-Gée, D. Lecomte and B. Ladevie. 2003. Drying kinetics and heat flux in thin layer conductive drying. Int. Com. on Heat and Mass Transfer, 30 (3) 335-349.

Maillet, D., S. Andre, J.C. Batsale , A. Degiovanni and C. Moyne. 2000. Thermal quadrupoles-Solving the heat eq. through int. transforms, John Wiley.

D. Petit, D. Maillet, *Techniques inverses et estimation de paramètres (Inverse techniques and parameter estimation),* Editeur : Techniques de l'Ingénieur, Paris. Thème : Sciences Fondamentales, base : Physique-Chimie, rubrique : Mathématiques pour la physique.
- Dossier AF 4515, pp. 1- 18, janvier 2008.
- Dossier AF 4516, pp. 1-24, janvier 2008.

Appendix 1

```
%Calculations related to the figure 2
% Deconvolution influence of the noise
% fi=fi0*exp(-((t-t0)/tau)*2)
% dT/dt = q - hT
% T=conv(q,exp(-ht))
% Tr = T + noise

N=100;dt=0.01;t0=dt*N/2;t=dt*(1:N);
q0=10;h=1;q=q0*exp(-20*(t-t0).^2);

% noise
std1=0.01;, std2=0.1;
noise1=std1*randn(size(t));noise2=std2*randn(size(t));

Imp=exp(-h*(t(1:N)));%Impulse response
X=dt*toeplitz(Imp, zeros(1,N)); % Sensitivity matrix

T=X*q'; % Direct model (convolution)
Tr1=T'+noise1;Tr2=T'+noise2;% noise simulation
G=inv(X'*X);
qr1=G*X'*Tr1'; qr2=G*X'*Tr2'; %OLS inversion

subplot(3,2,1), plot(t,Imp),title('Impulse response');
subplot(3,2,2), plot(t,q),title('Imput q(t)');
subplot(3,2,3), plot(t,T,t,Tr1), title(['Noisy signal std=' num2str(std1) 'k'])
subplot(3,2,4), plot(t,q,t,qr1,'*'), title(['Retrieved Input std=' num2str(std1) 'k'])
subplot(3,2,5), plot(t,T,t,Tr2), title(['Noisy signal std=' num2str(std2) 'k'])
subplot(3,2,6), plot(t,q,t,qr2,'*'), title(['Retrieved Input std=' num2str(std1) 'k'])
```

Appendix 2

```
%Calculations related to the figure 3
% Deconvolution with the TSVD method

N=100;dt=0.01;t0=dt*N/2;t=dt*(1:N);
q0=10;h=1;q=q0*exp(-20*(t-t0).^2);

% noise
std=0.1;
noise=std*randn(size(t));

Imp=exp(-h*(t(1:N)));%Impulse response
S=dt*toeplitz(Imp, zeros(1,N)); % Sensitivity matrix

T=S*q'; % Direct model (convolution)
Tr=T'+noise;% noise simulation

%TSVD

alph=[1 5 20 100];

for p=1:4
```

```
  [U,W,V]=svds(S,alph(p));
   qr=V*diag(1./diag(W))*U'*Tr';
 %  J(p)=norm(Tr-S*qr);
  % K(p)=norm(qr);

subplot(2,2,p),plot(t,T,'k',t,Tr,'k',t,q,'k:',t,qr,'k+'),
   title(['alpha = ',num2str(alph(p))])
   figure(gcf);
end
```

Appendix 3

```
%Calculations related to the figure 4
% Deconvolution with the Thikonnov method

N=100;dt=0.01;t0=dt*N/2;t=dt*(1:N);
q0=10;h=1;q=q0*exp(-20*(t-t0).^2);

% noise
std=0.1; %standard deviation of the noise
noise=std*randn(size(t));

Imp=exp(-h*(t(1:N)));%Impulse response
S=dt*toeplitz(Imp, zeros(1,N)); % Sensitivity matrix

T=S*q'; % Direct model (convolution)
Tr=T'+noise;% noise simulation

%TSVD

mu=[0.04 0.01 0.0007 0.000001];

for p=1:4

  %Thikonov regularization
   G=inv(S'*S+mu(p)*eye(N));
    qr=G*S'*Tr';

 %  J(p)=norm(Tr-S*qr);
 %  K(p)=norm(qr);

subplot(2,2,p),plot(t,T,'k',t,Tr,'k',t,q,'k:',t,qr,'k+'),
   title(['mu = ',num2str(mu(p))])
   figure(gcf);
end
```

Appendix 4

```
% Deconvolution and inversion with regularization related to figure 6
%L-Curve and comparisons netween TSVD and Thikonov regularization methods

clear
```

```
N=100;dt=0.01;t0=dt*N/2;t=dt*(1:N);
q0=10;h=1;q=q0*exp(-20*(t-t0).^2);

% noise
std=0.1;
noise=std*randn(size(t));

mu=[ 0.04 0.03 0.02 0.01 0.005 0.002 0.0015 0.001 0.0008 0.0005 0.0001 0.00005 5e-10 ];%
    Regularization parameter
nu=[1 2 3 4 5 6 10 20 30 40 50 70 100];%Truncation parameter

S=dt*toeplitz(exp(-h*(t(1:N))), zeros(1,N)); % Sensitivity matrix
T=S*q'; % Direct model (convolution)
Tr=T'+noise;

for i=1:length(mu)
%Thikonov regularization
  G=inv(S'*S+mu(i)*eye(N));
   qr=G*S'*Tr';
%TSVD regularization
  [U,W,V]=svds(S,nu(i));
  qrs=V*diag(1./diag(W))*U'*Tr';
 %norms
  nres(i)=norm(Tr-S*qr);
  nqr(i)=norm(qr);
  nress(i)=norm(Tr-S*qrs);
  nqrs(i)=norm(qrs);
end

hold off
plot(nres,nqr),xlabel('norm(T-S*q)'), ylabel('norm(q)'), hold on,plot(nress,nqrs,'o'),  figure 2
```

Lecture 7. Types of inverse problems, model reduction, model identification.

Part A: Experimental identification of low order model

Jean-Luc Battaglia

Email: jean-luc.battaglia@u-bordeaux.fr

University of Bordeaux
Laboratory I2M, Department TREFLE, UMR 5295
ENSAM, Esplanade des arts et métiers
33405 Talence Cedex, France.

Abstract: The system identification technique is used to formulate a reliable direct model to be used in an inverse heat transfer problem. This approach finds several practical applications in thermal sciences for reasons that will be developed in the text. For clarity, we will restrict our presentation to monovariable linear systems relating the temperature at one point in the system to one heat flux acting on the system. Two approaches are presented in this course. In the first one, the non-parametric method only uses the temperature and heat flux measurement by calculating the cross correlation or power spectral density. The second set of methods relates to the parametric methods that consist in identifying the parameters of a model that expresses the successive time derivatives of the temperature to the heat flux.

Nomenclature

a	Thermal diffusivity $m^2.s^{-1}$	S_{xy}	power spectral density between x and y
C_{xy}	correlation function between x and y	T	temperature, K
C_p	specific heat, $J.kg^{-1}.K^{-1}$	T	time, s
D^ν	derivative of real order ν	$X_s=[x_s,y_s]$	sensor coordinates
e	measurement error	y	temperature measurement, K
h_m	impulse response	V	loss function
h	exchange coefficient, $W.m^{-2}.K^{-1}$	Δt	sampling time
H	transfer function	φ	heat flux density $W.m^{-2}$

I^{ν}	integral of real order ν	ρ	density, kg m^{-3}
k	thermal conductivity, $W.m^{-1}.K^{-1}$		

Outline

1 Introduction

The system identification framework is a well-known domain that has applications in automatic (for control purpose mainly) and in signal processing [1][2]. For several years the heat transfer scientific community found very interesting applications of those methods for the modelling of heat and mass processes that occur in thermal systems [6][7][8]. In this course we present the system identification technique as an efficient tool to formulate a reliable direct model that can be used to solve the corresponding inverse heat transfer problem. In case of a *monovariable system*, as that represented in Figure 1, the inverse procedure will consist in estimating the heat flux acting on the studied system from temperature measurement at one point in the system. Let us highlight now that the methods that will be presented below can be obviously generalized to multivariable systems (several heat flux or heat sources acting on a system equipped with several sensors). As an additional constraint, we will also restrict the presentation of the methods to *linear systems*. It means that the thermal properties of the system will not depend on temperature. However, system identification has been developed for nonlinear systems, but mathematical derivations of such techniques are largely beyond the scope of this course.

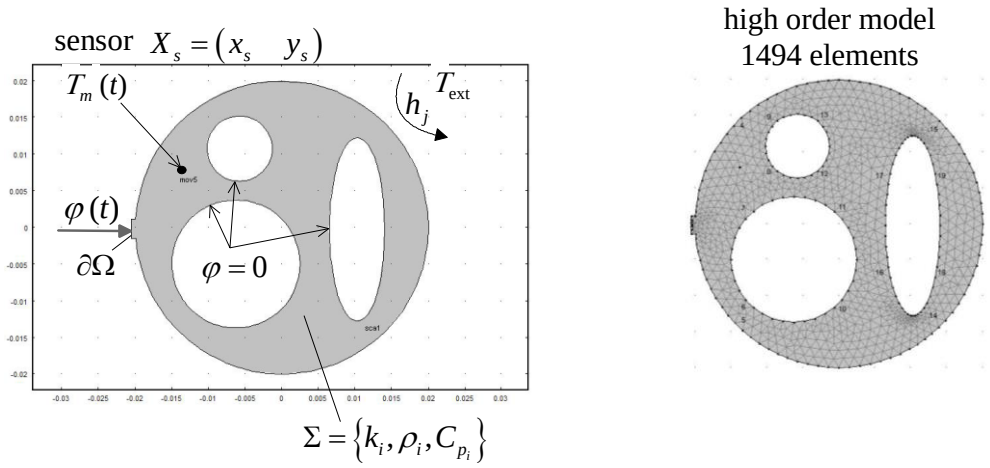


Figure 1: example of a 2D monovariable linear system.

Why are scientists working in the field of heat transfer, and more particularly in measurements inversion, interested with system identification? The first answer relates to model reduction. Indeed, whatever the implemented inverse technique, inversion requires using a direct model in an iterative manner to approach the solution. Statistical methods such as the Bayesian technique require calls upon the direct model many times and computational times could become dramatically long. As an example, let us consider the 2D system represented in Figure 1. The domain Σ is characterized by its thermal properties (thermal conductivity k_i, specific heat per unit volume $C_{p,i}$ and density ρ_i for constitutive material i). A heat flux φ is imposed on the boundary $\partial\Omega$ whereas the remaining part of the outdoor boundary is subjected to convection with the coefficient h_j, the temperature of the surrounding fluid being denoted T_{ext}. Finally, the inner boundaries are insulated. The objective here is to estimate the heat flux density

from temperature measurements in the plate. It is thus assumed that a sensor has been embedded in the plate and the temperature of the sensor is denoted $T_m(t)$. Although this problem is quite simple, only a discrete method (finite elements for example) can be used to solve the heat diffusion equation and associate boundary and initial conditions to simulate the temperature of the sensor. A mesh is thus built (see Figure 1) that leads to calculating the temperature at each node. This discrete model is so-called a *high-order model*, the order referring to the number of degrees of freedom of the mesh. Simulating the output of this model leads to results as those presented in Figure 2.

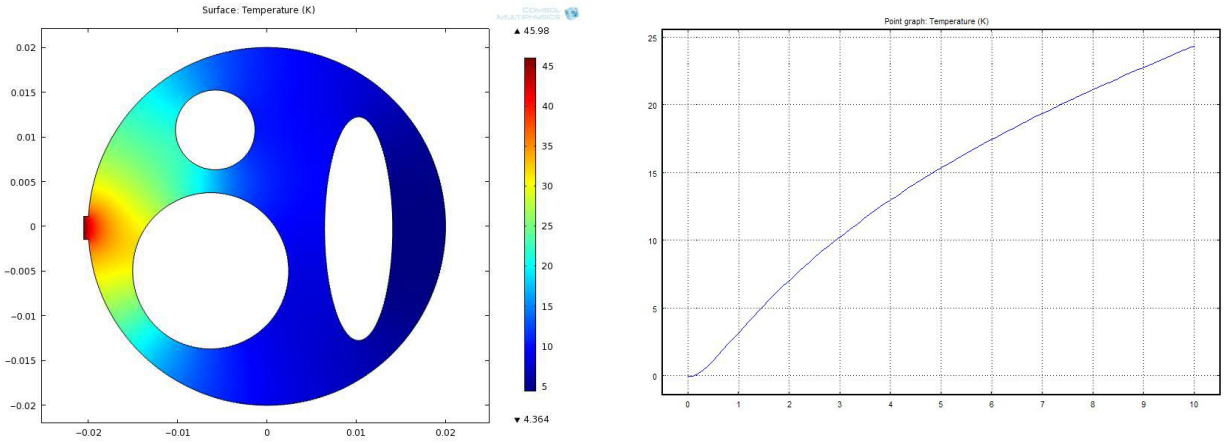


Figure 2: simulation of the temperature field at *t*=10 sec and of the time dependent temperature of the sensor for a step heat flux density.

The reliability of the direct model rests on the accuracy on two sets of data: the thermal properties $\{k_i, C_{p,i}, \rho_i, h_j\}$ and the location $X_s = [x_s, y_s]$ of the sensor. Uncertainties for these data will lead to a very low confidence domain for the estimated heat flux [9].
This system identification approach is described in a schematic way in Figure 3. The goal is to apply a known heat flux $\varphi(t)$ on the system and to measure the signal at the thermal sensor. We must note as a first point that *calibrating the sensor* (the link between the measured signal and the absolute temperature) is not required since the same sensor is used both for identifying the system and the above defined inversion. Once these data given, estimating "a" model **M** that relates them becomes possible. However, it must be emphasized that this estimated model has significance on the measurement time-domain only. *Prediction* is therefore a main issue of system identification. Secondly, the measurements are affected by an error (noise) that will have an influence on the identified model. It is generally admitted that the imposed heat flux is generally fully known and that it is errorless. Thus, all the error is reported on the sensor signal. **Obviously, the objective is to use a model M that is more accurate than that obtained from the FEM with uncertainties on $\{k_i, C_{p,i}, \rho_i, h_j\}$ and $X_s = [x_s, y_s]$.**

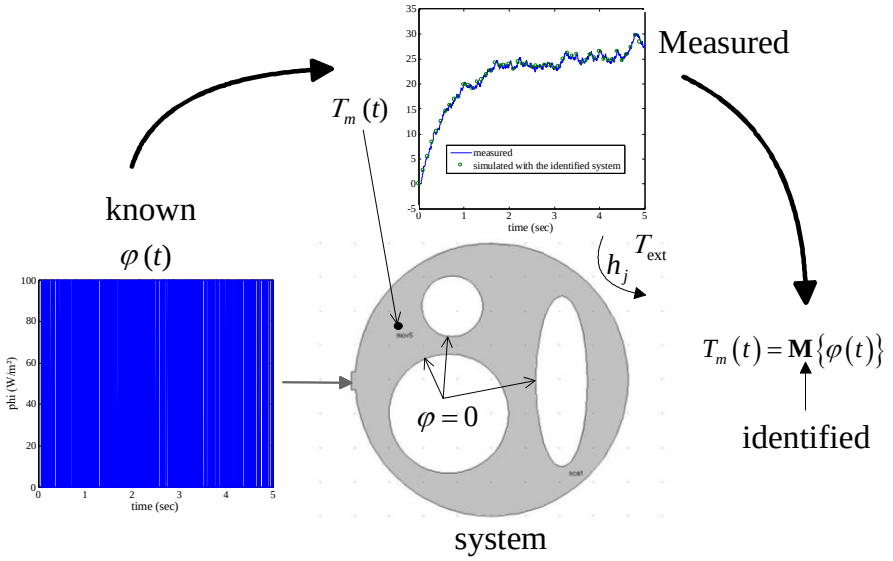


Figure 3: thermal system identification procedure.

Once the thermal system has been identified, it can be used to solve the inverse problem, which is to estimate the heat flux from model **M** and from temperature measurement at the sensors. The classical procedure is described in Figure 4.

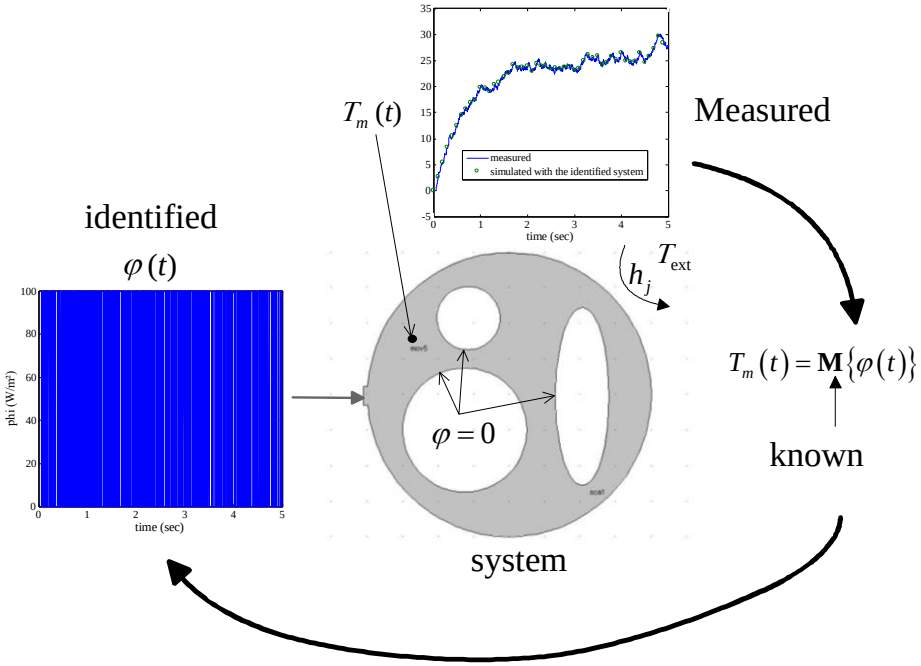


Figure 4: use of the identified system to solve the inverse procedure (estimating the heat flux).

It means that if the identified system model describes the thermal behaviour for the heat flux sequence represented in Figure 3 in a correct manner, it is then expected to retrieve this sequence applying an inverse technique starting from the knowledge of the identified model **M** and of the temperature measurement presented in Figure 3. This is what suggests Figure 4.

According to our previous description, it can be thus possible now to draw the main advantages and drawbacks of this approach.

Advantages

- The system identification approach will be first interesting to obtain a reliable and accurate *low order model* that will require less computational time for simulation.
- There is no need to know the thermal properties of the system (thermal conductivity, density, specific heat, heat exchange coefficients, thermal resistances at the interfaces, parameters related to thermal radiation…).
- It is not required to know the sensor location inside the system.
- Calibrating the sensor is not required.
- The identification procedure is fast (this will be viewed later with the description of the different techniques).

Drawbacks

- The model identification must be achieved in the same conditions as those encountered during the inversion (heat exchanges between the surrounding and the system must remain the same for the two configurations).
- The prediction of the identified model rests on strong assumptions (in particular, it is better reaching the stationary behaviour during the system identification process). In general, the identified system is only valid for the time duration of the system identification process.

2 The system identification approach

2.1 The impulse response

The temperature $T_m(t)$ of the sensor is related to the heat flux $\varphi(t)$ thanks to the impulse response $h_m(t)$ under the form of the following convolution product, that is a direct mathematical formulation of the Duhamel's theorem:

$$T_m(t) = T_m(0) + (h_m * \varphi)(t) = \int_0^t h_m(t-\tau)\,\varphi(\tau)\,\mathrm{d}\tau \tag{1}$$

Let us note here that if temperature $T_m(t)$ is expressed in kelvin, and if heat flux $\varphi(t)$ is in W.m^{-2}, this means that the product $\varphi(t)\mathrm{d}t$ is in J.m^{-2}, and consequently the impulse response $h_m(t)$ is in K.m^2.J^{-1}.

Equation (1) is valid for linear systems with a single transient excitation, with time independent coefficients and zero initial state, where the impulse response fully characterizes the forced thermal behavior. Therefore, any kind of inverse strategy can be based on the direct model expressed in terms of the impulse response of the system. However, as we said in the first section, this impulse response will depend on the following quantities: $\{k_i, C_{p,i}, \rho_i, h_j\}$ and $X_s = [x_s, y_s]$. According to the uncertainty that affects those quantities, the user could imagine directly measuring the impulse response from an experiment. It will consist in replacing the heat flux on the real problem by a *known* photothermal excitation, delivered by a laser for example, and in measuring the temperature of the sensor (case of a pulsed heat flux). However, this approach is not reliable since the impulse response magnitude should be very low to preserve the linear behavior of the system. As an illustration the temperature of the sensor is

calculated in the configuration given above with $\varphi(t) = 10^6 \times exp(-t^2/\tau^2)$ where $\tau = 1\mu$sec is small enough to consider this excitation as a Dirac distribution. The simulation is presented in Figure 5. The maximum amplitude of the response is very low here. The above approximation implies an additional contribution to the measurement error.

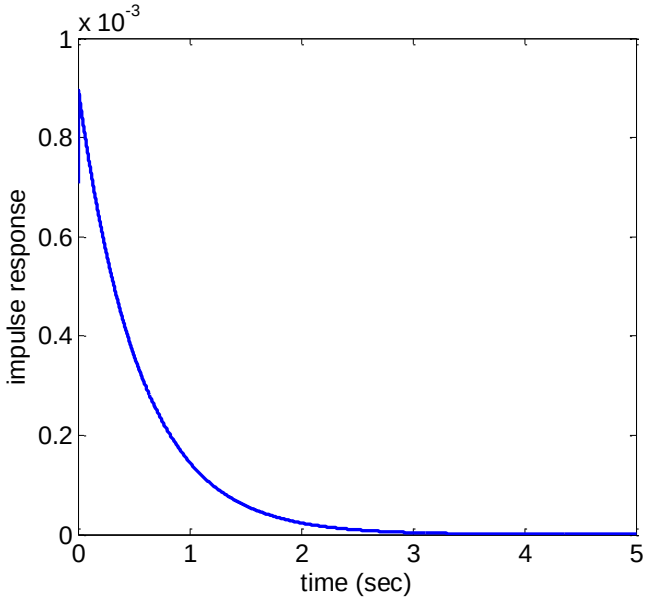


Figure 5: simulation of the impulse response using the FEM.

Another solution could consist in calculating the derivative of the step response shown in Figure 2 (right plot) to retrieve the impulse response. Again, it is not a reliable technique since the derivation will amplify the measurement error and will lead to a very inaccurate impulse response, especially at the short times.
Several powerful techniques have been developed in the system identification and signal processing domains that lead to more accurate identified impulse responses of the system. These techniques are classified in two sets of methods: the non-parametric methods and the parametric ones.

2.2 The non-parametric approach

2.2.1 The deconvolution technique

A very easy technique for the deconvolution of (1) is to consider the discrete form of this relation [2], obtained by the numerical quadrature of the convolution integral, calling $t_f = N\,\Delta t$ the duration of the experiment, where Δt is the sampling time interval:

$$T_k = T_m(k\,\Delta t) = T_m(0) + \sum_{i=1}^{k} \tilde{h}_{m,k-i+1}\,\tilde{\varphi}_i\,\Delta t \qquad (2)$$

$$\text{where } \begin{cases} \tilde{h}_{m,i} = \frac{1}{\Delta t}\int_{(i-1)\Delta t}^{i\Delta t} h_m(t)\mathrm{d}t \approx \frac{1}{2}\left(h_{m,i-1} + h_{m,i}\right) \text{ with } h_{m,i} = h_m(i\,\Delta t) \\ \tilde{\varphi}_i = \frac{1}{\Delta t}\int_{(i-1)\Delta t}^{i\Delta t} \varphi(t)\mathrm{d}t \approx \frac{1}{2}(\varphi_{i-1} + \varphi_i) \text{ with } \varphi_i = \varphi_i(i\,\Delta t) \end{cases} \tag{3}$$

Let us note that, in order to get a forced response, the three functions present in equation (1) should be equal to zero for times t such as $t \le 0$, and in particular $\varphi(0) = T_m(0) = 0 \Rightarrow h_m(0) = 0$, where the origin of time is the first time where flux φ departs from a zero value. So, the left-hand term of equation (2) is the instantaneous temperature at time $k\,\Delta t$, while the right terms correspond to average values of the impulse response and of the flux over a time interval. These average values are defined in equation (3) and correspond to the time interval $](i-1)\Delta t,\, i\,\Delta t]$ for $i \ge 1$.

Equation (2a) can be expressed under a vector/matrix form, calling $t_f = N\,\Delta t$ the duration of the experiment:

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ \vdots \\ T_N \end{bmatrix} = \begin{bmatrix} T_m(0) \\ T_m(0) \\ T_m(0) \\ \vdots \\ T_m(0) \end{bmatrix} + \Delta t \begin{bmatrix} \tilde{\varphi}_1 & & & & \\ \tilde{\varphi}_2 & \tilde{\varphi}_1 & & 0 & \\ \tilde{\varphi}_3 & \varphi_2 & \tilde{\varphi}_1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \tilde{\varphi}_N & \tilde{\varphi}_{N-1} & \tilde{\varphi}_{N-2} & \cdots & \tilde{\varphi}_1 \end{bmatrix} \begin{bmatrix} \tilde{h}_{m1} \\ \tilde{h}_{m2} \\ \tilde{h}_{m3} \\ \vdots \\ \tilde{h}_{mN} \end{bmatrix} \tag{4}$$

Assuming an additive measurement error of normal distribution (zero mean and constant standard deviation) for strictly positive times, with a zero error at initial time, the measured temperature is related to the exact one as:

$$y_m(k\,\Delta t) = T_m(k\,\Delta t) + e(k\,\Delta t) - T_m(0) = \Delta t \sum_{i=1}^{k} \tilde{h}_m(i\,\Delta t)\,\tilde{\varphi}((k-i+1)\,\Delta t) \tag{5}$$

Given that $\lim_{k\to\infty} h_k = 0$, it is reasonable to truncate the series from $k = Q$ and thus relation (5) becomes:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_Q \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{Y}_N} = \Delta t \underbrace{\begin{bmatrix} \tilde{\varphi}_1 & & & \\ \tilde{\varphi}_2 & \tilde{\varphi}_1 & 0 & \\ \vdots & & \ddots & \\ \tilde{\varphi}_Q & \cdots & \tilde{\varphi}_2 & \tilde{\varphi}_1 \\ \vdots & & & \vdots \\ \tilde{\varphi}_N & \cdots & \tilde{\varphi}_{N-Q+1} & \tilde{\varphi}_{N-Q+1} \end{bmatrix}}_{\mathbf{\Phi}_N} \underbrace{\begin{bmatrix} \tilde{h}_{m1} \\ \tilde{h}_{m2} \\ \vdots \\ \tilde{h}_Q \end{bmatrix}}_{\mathbf{H}_Q} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_Q \\ \vdots \\ e_N \end{bmatrix}}_{\mathbf{E}_N} \tag{6}$$

Vector H_Q can thus be estimated in the least square sense, to minimize $\left(\mathrm{E}_N\,\mathrm{E}_N{}^T\right)$ and one gets:

$$\mathrm{H}_Q = (\Phi_N\,\Phi_N^T)^{-1}\,\Phi_N^T\,\mathrm{Y}_N \tag{7}$$

However, this procedure is quite long according to the values of *Q* and *N* and very sensitive to measurement errors.

2.2.2 The correlation technique

A better and faster approach consists in identifying the impulse response $h(t)$, from the cross correlation product of the system response that is the temperature $T_m(t)$ of the sensor and the heat flux $\varphi(t)$ [1]. This method is well suited for non-causal systems, that is for problems where a space coordinate is the independent variable, or to systems where both excitation and response are time periodical. Indeed, let us rewrite relation (5) considering of the measurement errors and assuming that the heat flux is equal to zero for negative times as well as past the current time *t*:

$$y_m(t) = \int_0^t h_m(t-\tau)\,\varphi(\tau)\,d\tau + e(t) = \int_{-\infty}^{+\infty} h_m(t-\tau)\,\varphi(\tau)\,d\tau + e(t) \quad (8)$$
$$\text{where } \varphi(\tau) = 0 \text{ for } \tau \leq 0 \text{ and for } \tau > t$$

Now let us multiply the two members of this equality by the lagged heat flux $\varphi(t-\tau)$ and integrate~~s~~ from *t*=0 to infinity. We obtain then:

$$\int_{-\infty}^{+\infty} y_m(t)\varphi(t-\tau)\mathrm{d}\tau = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h_m(t-\tau)\varphi(\tau)\varphi(t-\tau)\mathrm{d}t\mathrm{d}\tau + \int_{-\infty}^{+\infty} \varphi(t-\tau)e(t)\mathrm{d}\tau \quad (9)$$

We see a convolution product between the two functions appears:

$$C_{y_m,\varphi} = \int_{-\infty}^{+\infty} h_m(t-\tau)\,C_{\varphi,\varphi}\mathrm{d}\tau + C_{e,\varphi} \quad (10)$$

If one chose the excitation sequence $\varphi(t)$ as a white noise:

$$C_{\varphi\varphi}(\tau) = \delta(\tau) \quad (11)$$

And finally, if one admits that the noise measurement is not correlated to the input signal ($C_{e,\varphi} = 0$), one has:

$$C_{y_m,\varphi}(\tau) = h(\tau) \quad (12)$$

It thus appears that the impulse response can be directly deducted from the correlation function between the temperature of the sensor and the heat flux. In practice the correlations functions are calculated using the Fast Fourier Transform of the signals.
The interest of the correlation analysis is to make the identification of the physical system under less energy constraints for the magnitude of the heat flux. Indeed, in opposition to pulse analysis, the energy does not have to be deposited in an intense way during a very short time (closest to a Dirac distribution). An interesting feature of such an approach is that the linearity and stationarity assumptions are clearly satisfied, and that the confidence domain of the estimated impulse response is the same ~~all~~ over all the explored frequency range.

2.2.3 Spectral technique

The correlation technique is very sensitive to the magnitude of the measurement noise and practically using the power spectral density instead of the correlation functions [4] is more interesting:

$$\mathrm{FFT}\left[C_{y_m\varphi}(\tau)\right] = \mathrm{FFT}\left[\int_0^\infty h_m(t-\tau)\, C_{\varphi\varphi}(\tau)\, \mathrm{d}\tau\right] = Y_m(f)\, \Phi(f) = S_{y_m\varphi}(f) \qquad (13)$$

and

$$\mathrm{FFT}\left[C_{\varphi\varphi}(\tau)\right] = \mathrm{FFT}\left[\int_0^\infty \varphi(t-\tau)\, \varphi(\tau)\, \mathrm{d}\tau\right] = \Phi(f)^2 = S_{\varphi\varphi}(f) \qquad (14)$$

$Y_m(f)$ and $\Phi(f)$ are the Fourier transforms of the temperature and of the heat flux respectively. In a similar way, $S_{\varphi\varphi}(f)$ and $S_{y_m\varphi}(f)$ are the auto and cross PSD (Power Spectral Density). Then, applying the Fourier transform on relation (10) yields immediately:

$$S_{y_m\varphi}(f) = H(f)\, S_{\varphi\varphi}(f) + S_{\varphi e}(f) \qquad (15)$$

Finally, assuming that the noise measurement is not correlated with the heat flux ($S_{\varphi e}(f) = 0$), the expression of the transfer function is:

$$H(f) = \frac{S_{y_m\varphi}(f)}{S_{\varphi\varphi}(f)} \qquad (16)$$

Since the duration of the experiment is set to a fixed value τ, the real input signal is:

$$\varphi_\Pi(t) = \varphi(t)\, \Pi_\tau(t) \qquad (17)$$

In this relation, $\Pi_\tau(t) = 1$ when $0 \le t \le \tau$ and 0 elsewhere. Then applying the Fourier transform for the heat flux leads to:

$$\Phi_\Pi(f) = \Phi(f) * \left(\tau \frac{\sin(\pi\, \tau\, f)}{\pi\, \tau\, f}\right) \qquad (18)$$

It appears that the Fourier transform of the heat flux is convoluted by the cardinal Sine function. Usually, the heat flux is pre-windowed by a specific function $g_\tau(t)$ that decreases the influence of the function $\Pi_\tau(t)$ as:

$$\varphi_\Pi(t) = \varphi(t)\, g_\tau(t) \qquad (19)$$

For example, the Hanning window [3][4] is often used. It is defined as:

$$g_\tau(t) = 0.5\left(1 - \cos\left(\frac{2\,\pi\, t}{\tau}\right)\right) \qquad (20)$$

An improved estimation of $S_{y_m\varphi}(f)$ and $S_{\varphi\varphi}(f)$ has also been proposed by Welch [5]. The method consists in dividing the time series data into possible overlapping segments, computing the auto and cross power spectral densities and averaging the estimates.

2.3 The parametric approach

2.3.1 Principle

The principles of the system identification method are presented by Ljung [1]. Assuming a linear and stationary system, that means that the thermal properties of the system do not vary with temperature and time, the method consists in identifying the parameters involved in a linear relation between the heat flux $\varphi(t)$ and the temperature $T_m(t)$ of the sensor, from

measurements of these two quantities. Without any kind of physical consideration of the heat transfer process, it is assumed a general relationship of the following form is assumed:

$$T_m(t) + \alpha_1 \frac{\mathrm{d}T_m(t)}{\mathrm{d}t} + \alpha_2 \frac{\mathrm{d}^2 T_m(t)}{\mathrm{d}t^2} + \cdots = \beta_0 \varphi(t) + \beta_1 \frac{\mathrm{d}\varphi(t)}{\mathrm{d}t} + \beta_2 \frac{\mathrm{d}^2 \varphi(t)}{\mathrm{d}t^2} + \cdots \quad (21)$$

This kind of model is consistent with the behaviour of the dynamical systems, and it is also the case for thermal systems since the heat diffusion equation rests on the first order derivative of the temperature for all the points of the system. It is thus reasonable to admit that the temperature at time *t* must depend on the heat flux value at time *t* and at previous times. On the other hand, since temperature at times before *t* depends on the heat flux at previous times also, it is not surprising that they appear in the model.

Let us illustrate it on a simple configuration by considering the one-dimensional heat transfer in a wall (thermal conductivity *k* and thermal diffusivity *a*) subjected to the heat flux $\varphi(t)$ at $x = 0$ and insulated on the other face at $x = e$. The model is thus:

$$\frac{\partial T(x,t)}{\partial t} = a \frac{\partial^2 T(x,t)}{\partial x^2}, 0 < x < e, t > 0 \quad (22)$$

The boundary conditions are:

$$-k \frac{\partial T(x,t)}{\partial x} = \varphi(t), x = 0, t > 0 \quad (23)$$

$$\frac{\partial T(x,t)}{\partial x} = 0, x = e, t > 0 \quad (24)$$

And the initial condition is chosen as:

$$T(x,t) = 0, 0 \le x \le e, t = 0 \quad (25)$$

Let us examine the temperature at $x = e$ and we note $T_m(t) = T(x = e, t)$. The Laplace transform $L\{\ \}$ is used to solve the previous problem:

$$L\{T_m(t)\} = \theta_m(s) = \frac{1}{k\,\beta\,\sinh(\beta\, e)} L\{\varphi(t)\} = \frac{1}{k\,\beta\,\sinh(\beta\, e)} \Phi(s) \quad (26)$$

Where: $\beta = \sqrt{s/a}$. The hyperbolic sine function can be expressed as the following series:

$$\sinh(z) = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!}, \forall z \ge 0 \quad (27)$$

Replacing this expression in relation (26) yields:

$$\theta_m(s) = \frac{1}{k\,\beta \sum_{n=0}^{\infty} \frac{(\beta e)^{2n+1}}{(2n+1)!}} \Phi(s) = \frac{1}{k \sum_{n=0}^{\infty} \frac{e^{2n+1} s^{n+1}}{a^{n+1}(2n+1)!}} \Phi(s) \quad (28)$$

That can be also written as:

$$\sum_{n=0}^{\infty} \alpha_n s^{n+1} \theta_m(s) = \Phi(s) \quad (29)$$

With: $\alpha_n = k \frac{e^{2n+1}}{a^{n+1}(2n+1)!}$.

At this stage we must remind us of an important property related to the Laplace transform of the derivative of a function:

$$L\left(\frac{\mathrm{d}^n f(t)}{\mathrm{d}t^n}\right) = s^n F(s) - \sum_{k=0}^{n-1} s^{n-k-1} \frac{\mathrm{d}^k f(0)}{\mathrm{d}t^k} \tag{30}$$

Taking into account the initial condition (25), relation (29) becomes :

$$\sum_{n=0}^{\infty} \alpha_n \frac{\mathrm{d}^{n+1} T_m(t)}{\mathrm{d}t} = \varphi(t) \tag{31}$$

It is therefore demonstrated that the heat transfer model expressing the temperature at $x = e$ according to the heat flux $\varphi(t)$ imposed at $x = 0$ can be put on the form of ~~the~~ relation (21). In fact the series in (31) can be significantly truncated and we will thus obtain a *low order model*.

The discrete form of the derivatives gives rise to an equivalent form of relation (21) and temperature at time $k\Delta t$ depends on the heat flux and the temperature at previous times as:

$$T_m(k) = b_0\varphi(k) + b_1\varphi(k-1) + b_2\varphi(k-2) + \cdots - a_1 T_m(k-1) - a_2 T_m(k-2) - \cdots \tag{32}$$

Let us note that replacing the temperature at previous times with the measurement in relation (32)leads to the predictive model as:

$$\hat{T}_m(k) = b_0\varphi(k) + b_1\varphi(k-1) + b_2\varphi(k-2) + \cdots - a_1 y_m(k-1) - a_2 y_m(k-2) - \cdots \tag{33}$$

Relation (32) is called the output error model whereas relation (33) is called the predictive model. Identification of parameters (a_i, b_j) will significantly differ according to the choice of the model as represented in Figure 6.

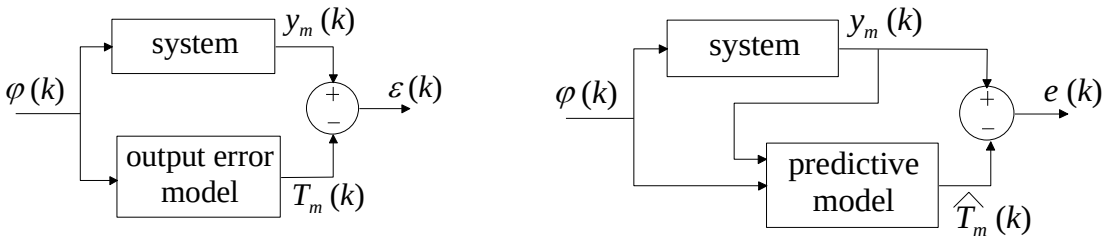


Figure 6: parameter identification according to the model representation (output error or predictive).

In case of the output error model configuration, the sensitivity functions $S(a_i) = \mathrm{d}T_m(t)/\mathrm{d}a_i$ and $S(b_j) = \mathrm{d}T_m(t)/\mathrm{d}b_j$ depend on the parameters a_i and b_j. It means that the minimization of $\rho(N) = \sum_{k=0}^{N} \varepsilon(k)^2$ requires a non-linear minimization algorithm. On the other side, the sensitivity functions do not depend anymore on the parameters when minimizing the quantity

$r(N)=\sum_{k=0}^{N} e(k)^2$. It means that estimation of the parameters in case of the predictive model is implemented by a linear minimization algorithm.

2.3.2 Output error model

Let us assume that the number of parameters is *n* for a_i and (*n*+1) for b_j. The sensitivity functions of the temperature at time $k\Delta t$ with respect to a_i and b_j are:

$$S_{a_i}(k) = \frac{\partial T_m(k)}{\partial a_i}, i = 1, \dots, n \tag{34}$$

$$S_{b_i}(k) = \frac{\partial T_m(k)}{\partial b_i}, i = 0, \dots, n \tag{35}$$

According to relation (32), it is obtained:

$$S_{a_i}(k) + a_1 S_{a_i}(k-1) + \cdots + a_n S_{a_i}(k-n) = -T_m(k-i), i = 1, \dots, n \tag{36}$$

With: $S_{a_i}(0) = S_{a_i}(1) = \cdots = S_{a_i}(n-1) = 0$
And:

$$b_0 S_{b_i}(k) + b_1 S_{b_i}(k-1) + \cdots + b_n S_{b_i}(k-n) = \varphi(k-i), i = 0, \dots, n \tag{37}$$

With: $S_{b_i}(0) = S_{b_i}(1) = \cdots = S_{b_i}(n-1) = 0.$
Therefore, the output error at time $k\Delta t$ is:

$$\varepsilon(k) = y_m(k) - T_m(k) = \sum_{i=1}^{n} S_{a_i}(k)\Delta a_i + \sum_{i=0}^{n} S_{b_i}(k)\Delta b_i \tag{38}$$

Let us imagine that measurements are collected from $n\Delta t$ up to $N\Delta t$. A matrix representation of (38) of the following form is thus obtained:

$$\mathbf{E} = \begin{bmatrix} \varepsilon(n) \\ \varepsilon(n+1) \\ \vdots \\ \varepsilon(N) \end{bmatrix} = \mathbf{S} \begin{bmatrix} \Delta a_1 \\ \vdots \\ \Delta a_n \\ \Delta b_0 \\ \vdots \\ \Delta b_n \end{bmatrix} = \mathbf{S}\Delta\Theta \tag{39}$$

Where:

$$\mathbf{S} = \begin{bmatrix} S_{a_1}(n) & \cdots & S_{a_n}(n) & S_{b_0}(n) & \cdots & S_{b_n}(n) \\ \vdots & & \vdots & \vdots & & \vdots \\ S_{a_1}(N) & \cdots & S_{a_n}(N) & S_{b_0}(N) & \cdots & S_{b_n}(N) \end{bmatrix} \tag{40}$$

Solving equation (39) in the least square sense leads to:

$$\Delta\Theta = (\mathbf{S}^T\,\mathbf{S})^{-1}\mathbf{S}^T\,\mathbf{E} \tag{41}$$

It is thus possible to obtain the optimal value of Θ using an iterative scheme as:

$$\Theta_v = \Theta_{v-1} + \Delta\Theta_{v-1} \tag{42}$$

2.3.3 Predictive model

Relation (33) can be put under the form:

$$y_m(k) = \mathbf{H}(k)\,\Theta + e(k) \tag{43}$$

Where $\Theta^T = [a_1 \quad \cdots \quad a_n \quad b_0 \quad \cdots \quad b_n]$ and $\mathbf{H}$ is the regression vector defined as:

$$\mathbf{H}(k) = [-y_m(k-1) \quad \cdots \quad -y_m(k-n) \quad \varphi(k) \quad \cdots \quad \varphi(k-n)] \tag{44}$$

Let us imagine that measurements are collected from $n\Delta t$ up to $N\Delta t$. Therefore, relation (43) leads to:

$$\mathbf{Y}_N = \Psi_N \Theta + \mathbf{E}_N \tag{45}$$

Where:

$\mathbf{Y}_N^T = [y_m(n) \quad \cdots \quad y_m(N+n)]$, $\Psi_N^T = [\mathbf{H}(n) \quad \cdots \quad \mathbf{H}(N+n)]$ and $\mathbf{E}_N^T = [e(n) \quad \cdots \quad e(N+n)]$.

An estimation of Θ in the linear least square sense is obtained ~~as~~:

$$\widehat{\Theta} = (\Psi_N\,\Psi_N^T)^{-1}\Psi_N{}^T\,\mathbf{Y}_N \tag{46}$$

Despite ~~of~~ the rapidity of the method, it must be noted that the estimation is biased. Indeed, let us replace the expression of the identified parameters, relation (46), in the model, relation (43) . It is found:

$$\widehat{\Theta} = \Theta + (\Psi_N\,\Psi_N^T)^{-1}\,\Psi_N^T\,\mathbf{E}_N \tag{47}$$

It is demonstrated in the literature that:

$$E\{\widehat{\Theta}\} = \Theta + (E\{\mathbf{H}(k)\,\mathbf{H}(k)^T\})^{-1}E\{\mathbf{H}(k)^T\,e(k)\} \tag{48}$$

It thus appears that if $e(k)$ is correlated with $\mathbf{H}(k)$ or if $E\{e(k)\}$ is not zero, the estimation is biased and $E\left\{\widehat{\Theta}\right\} \neq \Theta$.

To accelerate the identification of Θ, a recursive scheme can be used. The vector of parameters at instant *t* is estimated from parameters estimated previously at instant $(t-1)$ according to:

$$\widehat{\Theta}(k) = \widehat{\Theta}(k-1) + \mathbf{L}(k)\left[y_m(k) - \mathbf{H}(k)\widehat{\Theta}(k-1)\right] \tag{49}$$

With:

$$\mathbf{L}(k) = \frac{\mathbf{P}(k-1)\,\mathbf{H}(k)^T}{\lambda(k)+\mathbf{H}(k)\;\mathbf{P}(k-1)\,\mathbf{H}(k)^T}$$

And:

$$\mathbf{P}(k) = \mathbf{P}(k-1) - \frac{\mathbf{P}(k-1)\mathbf{H}(k)^T\mathbf{H}(k)\mathbf{P}(k-1)}{\lambda(k)+\mathbf{H}(k)\mathbf{P}(k-1)\mathbf{H}(k)^T}$$

where the initial values are: $\widehat{\Theta}(0) = \mathbf{0}_D$ and $\mathbf{P}(0) = 10^6\mathbf{I}_D$, with $\mathbf{0}_D$ and $\mathbf{I}_D$ are zeros vector and unity matrix respectively with dimension $D = 2N$.

Remark: unbiased approaches are proposed in the literature that consist in whitening the sequence $e(k)$ in relation to (43). This is the instrumental variables method, and methods based on the change of the model structure (auto regressive with exogeneous input model = ARX, auto regressive with adjusted mean and exogeneous input model for example).

3 Input signal waveform – the PRBS signal

Whatever the method used, non-parametric or parametric, the choice of the input sequence is crucial regarding the quality of the identified system. In practice, we will consider the heat flux sequence as a Pseudo Random Binary Signal (PRBS). "White noise" is the term given to completely random unpredictable noise, such as the hiss you hear on an untuned radio. It has the property of having components at every frequency. A pseudo-random binary sequence (PRBS) can also have this property but is entirely predictable. A PRBS is rather like a long recurring decimal number- it looks random if you examine a short piece of the sequence, but it repeats itself every *m* bit. Of course, the larger *m* is, the more random it looks. You can generate a PRBS with a shift register and an XOR gate. Connecting the outputs of two stages of the shift register to the XOR gate, and then feeding the result back into the input of the shift register will generate a PRBS of some sort. Some combinations of outputs produce longer PRBSs than others- the longest ones are called m-sequences (where m means "maximum length"). A binary sequence (BS) is a sequence of *N* bits,

$$a_j \text{ for } j = 0, 1, ..., N-1$$

i.e. *m* ones and $N - m$ zeros. A BS is pseudo-random (PRBS) if its autocorrelation function:

$$C(v) = \sum_{j=0}^{N-1} a_j \, a_{j+v} \tag{50}$$

has only two values:

$$C(v) = \begin{cases} m, \text{ if } v \equiv 0 \pmod N \\ m \times c, \text{otherwise} \end{cases} \tag{51}$$

Where:

$$c = \frac{m-1}{N-1} \tag{52}$$

is called the duty cycle of the PRBS.
A PRBS is random in a sense that the value of an a_j element is independent of the values of any of the other elements, like real random sequences.
It is 'pseudo' because it is deterministic and after *N* elements it starts to repeat itself, unlike real random sequences, such as sequences generated by radioactive decay or by white noise. The PRBS is more general than the n-sequence, which is a special pseudo-random binary sequence of n bits generated as the output of a linear shift register. An n-sequence always has a 1/2 duty cycle and its number of elements $N = 2^k - 1$.

4 Application

Let us consider the heat transfer problem presented at the beginning and let us generate a heat flux sequence under the form of the pseudo random binary sequence represented in Figure 7. Such a choice for the excitation sequence makes the identification quite easy in practice. This sequence is also very close to a white noise in terms of the power spectral density ~~as~~ represented in Figure 8.

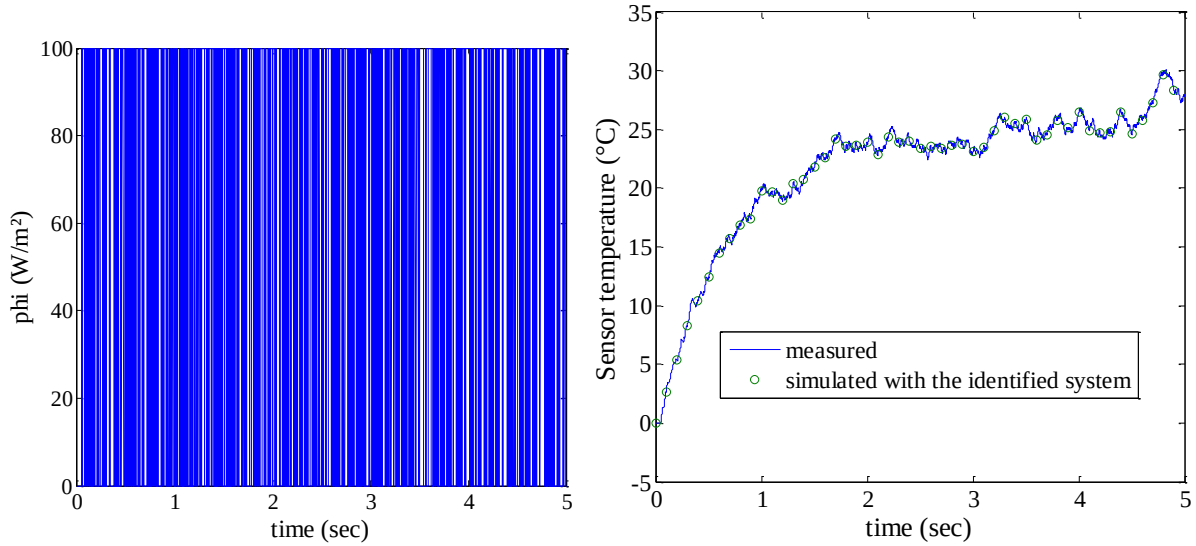


Figure 7: image on the left – heat flux generated on the form of a PRBS; image on the right – measured temperature of the sensor and comparison with the simulation of the identified system.

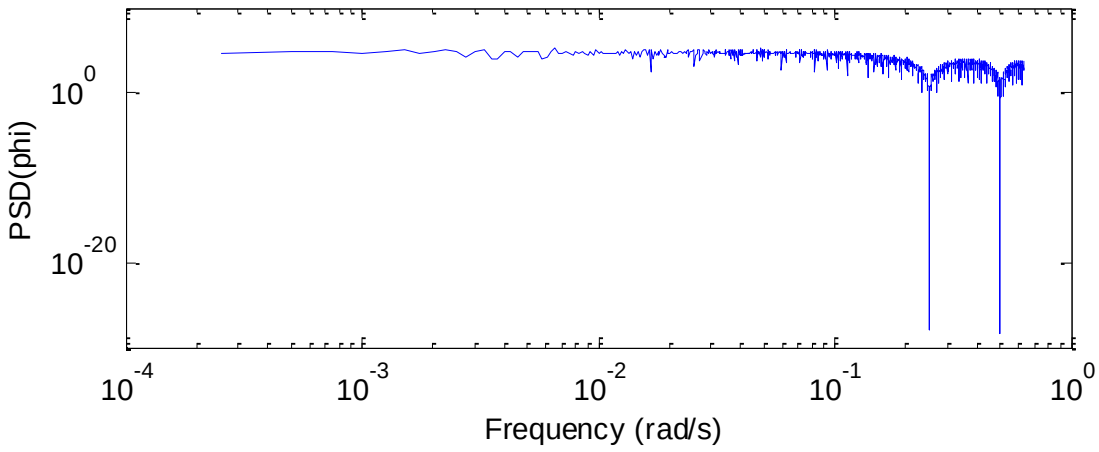


Figure 8: power spectral density of the heat flux generated as a PRBS.

Using the correlation method described previously, the impulse response represented in Figure 9 is obtained. This figure shows that the impulse response reconstructed using the correlation technique is very sensitive to the measurement noise.

In a second stage, we used the parametric approach in order to find the model on the form of the relation (33) that best fits the experimental measurements (Figure 7). The choice of $\mathbf{\Lambda} = [na, nb]$ (*na* is the number of parameters a_i and *nb* is the number of parameters b_i) is made by collecting in a matrix all the values of $\mathbf{\Lambda}$ to be investigated and looking for the value of the Aikake [1] criterion defined by

$$\Psi = \frac{1+n/N}{1-n/N} V, n = na + nb + 1 \tag{53}$$

where *n* is the total number of estimated parameters and *V* is the loss function defined by

$$V = \sum_{k=1}^{N} e_k^2 \tag{54}$$

Standard errors of the estimates are calculated from the covariance matrix of $\widehat{\Theta}$. If the assumptions of additive, zeros mean, constant variance σ^2 and uncorrelated errors are verified, the covariance matrix is expressed as

$$\text{cov}\,(\widehat{\mathbf{\Theta}}) = (\mathbf{H}^T\,\mathbf{H})^{-1}\boldsymbol{\sigma}^2 \tag{55}$$

An estimate of the variance σ^2, denoted s^2, is:

$$s^2 = \frac{1}{N-n}\mathbf{E}^T\,\mathbf{E} \tag{56}$$

It is found the optimal set of parameters (a_i, b_i) as:

Parameter	value	Standard deviation	Parameter	value	Standard deviation
a_0	1	0	a_5	0.0166	0.0054
a_1	0.2823	0.01364	b_0	0.0007006	5.348e-006
a_2	0.2539	0.01368	b_1	0.0006788	1.19e-005
a_3	0.2715	0.01375	b_2	0.0004693	1.404e-005
a_4	0.2047	0.01427	b_3	0.0002561	1.365e-005

The loss function is *V*=0.000123859.

Simulating the response with the heat flux sequence brings a very good agreement with measured data as represented in Figure 7. Therefore, the impulse response of the identified system is reported in Figure 9. A very nice agreement with that calculated from the FEM is found. The main difference occurs for short times.

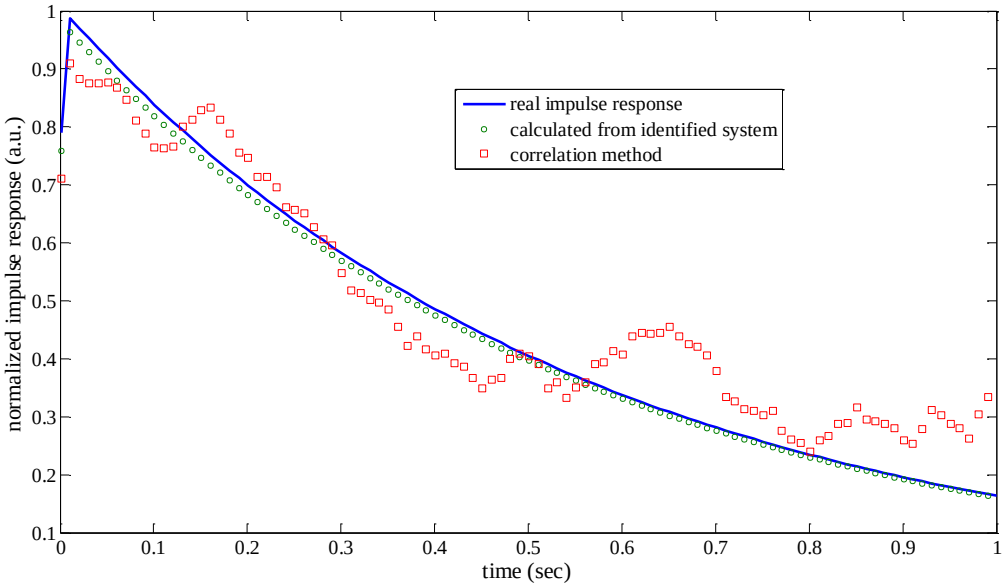


Figure 9: real impulse response and impulse response found using the correlation method and the parametric method.

5 Conclusion

System identification is a powerful tool that allows the user to obtain a direct model to solve an inverse problem. In fact, this approach consists in applying a known thermal excitation and to measure the temperature at the sensors to find a relationship between these two quantities. Obviously, this approach finds an interest if the system is not well characterized in terms of its thermal properties (thermal conductivity, specific heat, density, heat exchange coefficient at the boundaries, thermal resistance at the interfaces). Moreover, this technique does not require knowing the exact locations of the sensors in the system as well as their dynamical behaviour. It means that a calibration of the sensors is not required since they are used both for the system identification and the inversion. The constraints encountered with such an approach are that the system must be identified in the same configuration in which it will be used during the inversion. It means first that the time range for the system identification will define the usable time domain for the direct model. On the other hand, all the boundary conditions experienced during the system identification must remain identical during the inversion.
Finally, it must be emphasized than the computational times for the inversion will be decreased very significantly even if the thermal system is complex. It is a very interesting feature of this approach since the simulation of the identified system is faster than that based on a discretization of the heat equation.

6 References

[1] Ljung, L., *System identification. Theory for the user*, Prentice – Hall, inc., Englewood Cliffs, New Jersey, 1986.

[2] Walter, E., Pronzato, L., *Identification de modèles paramétriques à partir de données expérimentales*, Collection Modélisation Analyse Simulation et Commande, Éditions Masson, 1994.

[3] Kay S.M., *Modern Spectral Estimation*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[4] Stoïca P. and Moses R., *Introduction to spectral analysis*, Upper Saddle River, NJ: Prentice-Hall, 1997.

[5] Welch, P. D., *The use of Fast Fourier Transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms*, IEEE Trans. Audio Electroacoust., Vol. AU-15, 70-73, June 1967.

[6] Battaglia J.-L., Batsale J.-C., *Estimation of heat flux and temperature in a tool during turning*, Inverse Problems in Engineering, vol. 8, p.435-456,2000.

[7] Battaglia J.-L., Le Lay L.., Batsale J.-Ch. L., Oustaloup A., Cois O., *Utilisation de modèles d'identification non entiers pour la résolution de problèmes inverses en conduction*, International Journal of Thermal Science, Vol. 39, pp. 374-389, 2000.

[8] Battaglia J.-L., Cois O., Puigsegur L., Oustaloup A., *Solving an inverse heat conduction problem using a non-integer identified model*, Int. J. Heat Mass Trans., 44, 2001, pp. 2671-2680.

[9] Beck J., Blackwell B., St. Clair C.R., *Inverse Heat Conduction*, A Wiley-Interscience Publication, 1985.

[10] Maillet D., André S., Batsale J.-C., Degiovanni A. et Moyne C., *Thermal quadrupoles. Solving the heat equation through integral transforms*, ed. Wiley, 2000.

[11] Fourier J., *Théorie analytique de la chaleur*, Paris, 1822.

[12] Liouville J., *Mémoire sur quelques quéstions de géométrie et de mécanique, et sur un noveau genre pour réspondre ces quéstions*, Jour.Ecole Polytech. 13 (1832) 1-69.

[13] Oldham K. B., Spanier J., *The fractional calculus*, Academic Press, New York and London, 1974.

[14] Oldham K. B., Spanier J., *The replacement of Fick's laws by a formulation involving semi differentiation*, Electroanalytical Chemistry and Interfacial Electrochemistry 26 (1970) 331-341.

[15] Oldham K. B., Spanier J., *A general solution of the diffusive equation for semi-infinite geometries*, Journal of Mathematical Analysis and Applications 39 (1972) 655-669.

[16] Battaglia J.-L., Le Lay L., Batsale J.-C., Oustaloup A., Cois O., *Heat flow estimation through inverted non integer identification models*, International Journal of Thermal Science 39 (3) (2000) 374-389.

[17] Battaglia J.-L., Cois O., Puigsegur L., Oustaloup A., *Solving an inverse heat conduction problem using a non-integer identified model*, Int. J. of Heat and Mass Transfer 14 (44) (2000) 2671-2680.

Lecture 7: Types of inverse problems, model reduction, model identification.

Part B: Modal reduction for thermal problems: Core principles and presentation of the AROMM method

F. Joly, **Y. Rouizi**, **B. Gaume**, **O. Quéméner**

Laboratoire de Mécanique et d'Energétique d'Evry, Univ. Paris-Saclay
40 rue du Pelvoux, Courcouronnes 91020 Evry, France

E-mail: f.joly@iut.univ-evry.fr
yassine.rouizi@univ-evry.fr
b.gaume@iut.univ-evry.fr
o.quemener@iut.univ-evry.fr

Abstract. In the second part of this lecture, the special case of modal reduction is discussed. This method allows to greatly reduce the size of the model in case of complex geometry. The principle of this technique is presented. A focus on the AROMM method is carried out. We insist on the necessity to choose a modal basis adapted to the physical problem. The different principles of bases reduction are introduced.

Scope

1 Introduction

As computer hardware developing, the requirements in respect of numerical simulation follow the same pattern. They are therefore becoming more demanding.
First, one have to use geometries that perfectly match the reality of the simulated object. A recent study [4] has shown that the exact numerical modelling of a simple electronic component needs a mesh of $422k$ nodes. This order of magnitude has to be compared to industrial demand, that is to obtain the simulation of an entire electronic card.
Furthermore, we are also looking for being more and more precise taking into account physical phenomena. In thermal problems, infrared radiations for example, hugely complicate the heat transfer simulations [3].
Considering the inverse approach, this effect is amplified by the iterative procedure which involve the use of an important number of simulations [1].
For all those reasons, the use of reduced models is a topical issue. The idea consist in searching the temperature field as a whole by using a small number of unknowns.

2 Context of the study: the heat equation

The problem is the following: the domain Ω, delimited by boundary Γ, is characterized by its thermal conductivity $k(M,t)$ [W.m^{-1}.K^{-1}] and its volumetric heat capacity $c(M,t)$[J.m^{-3}.K^{-1}]. This domain receives two types of thermal loadings:

- the influence of the environment, which is characterised by a temperature $T_f(M,t)\,[K]$ and a heat exchange coefficient $h(M,t)$[W.m^{-2}.K^{-1}],
- the thermal dissipation, which can be a volumetric power on the domain $\pi(M,t)$[W.m^{-3}] or a surface load on the border $\varphi(M,t)$[W.m^{-2}].

Such a problem corresponds to the following equations:

$$\left\{\begin{array}{lcl} \forall M\in\Omega & : & c\dfrac{\partial T}{\partial t}=\overrightarrow{\nabla}.(k\overrightarrow{\nabla}T)+\pi \\ \forall M\in\Gamma & : & k\overrightarrow{\nabla}T.\overrightarrow{n}=\varphi+h(T_f-T) \end{array}\right. \tag{1}$$

For complex geometries, the solution of this problem is numerical and needs a spatial discretization. The finite element method leads to the weak variational formulation of (1). Let g be the test function, defined on the Hilbert space $H^1(\Omega)$, we can write:

$$\begin{array}{rl} \forall g\in H^1(\Omega),\quad \displaystyle\int_\Omega g\,c\,\frac{\partial T}{\partial t}\,d\Omega= & -\displaystyle\int_\Omega k\overrightarrow{\nabla}g.\overrightarrow{\nabla}T\,d\Omega-\int_\Gamma g\,h\,T\,d\Gamma \\ & +\displaystyle\int_\Omega g\,\pi\,d\Omega+\int_\Gamma g\,(\varphi+h\,T_f)\,d\Gamma \end{array} \tag{2}$$

It should be noted that it would be possible to consider:

- an anisotropic thermal conductivity characterized by a tensor $\overline{\overline{\mathbf{k}}}$,
- an advection - conduction problem, for which we add to the heat equation a transport term,
- infrared radiation between boundaries.

[1] In case of linear inverse problem, even if it is possible to use a direct procedure, this one needs one matrix inversion.

The addition of these terms does not change anything for the reduction method, and we will consider afterwards the problem defined by (1).
The spatial discretization of (2) leads to the following equation (according to the order of terms) :

$$\mathbf{C}\frac{d\mathbf{T}}{dt} = \mathbf{A}\mathbf{T} + \mathbf{U} \tag{3}$$

where $\mathbf{C}$ et $\mathbf{A}$ are respectively named the capacity matrix and the conductivity matrix, with a dimension $[N \times N]$, where N is the degrees of freedom (DOF) for the considered discretized domain. $\mathbf{T}$ is the temperature vector, which depends on the time, and $\mathbf{U}$ is the load vector. The dimension of all these vectors are $[N \times 1]$.
This equation constitutes the complete heat problem, which the DOF can be very important[2] in case of complex geometry.

3 The modal reduced model principle

This method is based on the time-space separation:

$$T(M,t) = \sum_{i=1}^{\infty} V_i(M)\, x_i(t) \tag{4}$$

Considering the space function $V_i(M)$ as being known, it means that the calculation of the temperature fields correspond to compute excitation states $x_i(t)$ of these functions. It is important to notice that the relation (4) is true only if the space functions $V_i(M)$ constitute a basis of the solutions space of the thermal problem (2), and this is not systematic.
The idea is then to rewrite this formulation using a limited number n of space functions $\widetilde{V}_i(M)$, which leads to an acceptable reconstitution of the thermal fields $\widetilde{T}(M,t) \simeq T(M,t)$:

$$\widetilde{T}(M,t) = \sum_{i=1}^{n} \widetilde{V}_i(M)\, \widetilde{x}_i(t) \tag{5}$$

Whatever the reduction technique used, the reduced model is obtained by projection of the heat equation on the subspace defined by the space functions $V_i(M)$. The equation (2) then becomes :

$$\begin{aligned} &\forall g \in H^1(\Omega), \\ &\int_\Omega g\, c\, \frac{\partial}{\partial t}\left(\sum_{i=1}^{n} \widetilde{V}_i\, \widetilde{x}_i\right) d\Omega = \\ &\qquad -\int_\Omega k \overrightarrow{\nabla} g . \overrightarrow{\nabla}\left(\sum_{i=1}^{n} \widetilde{V}_i\, \widetilde{x}_i\right) d\Omega - \int_\Gamma g\, h \left(\sum_{i=1}^{n} \widetilde{V}_i\, \widetilde{x}_i\right) d\Gamma \\ &\qquad +\int_\Omega g\, \pi\, d\Omega + \int_\Gamma g\, (\varphi + hT_f)\, d\Gamma \end{aligned} \tag{6}$$

In considering that all the space functions $\widetilde{V}_i(M)$ form a basis for the physical problem, these functions can be used as test functions for the variational formulation: $g(M) = \widetilde{V}_j(M)$.
After rearrangement, we have:

[2] For a finite volume method or for the finite element method for which the interpolation functions are linear, the DOF corresponds to the N number of mesh nodes.

$$\begin{aligned}
&\forall\ \widetilde{V}_j \in H^1(\Omega), j \in \mathbb{N},\\
&\quad \sum_{i=1}^{n} \left(\int_\Omega \widetilde{V}_j\ c\ \widetilde{V}_i\, d\Omega \right)\ \frac{d\widetilde{x}_i}{dt} =\\
&\qquad -\sum_{i=1}^{n} \left(\int_\Omega k \overrightarrow{\nabla} \widetilde{V}_j . \overrightarrow{\nabla} \widetilde{V}_i\, d\Omega + \int_\Gamma \widetilde{V}_j\, h\, \widetilde{V}_i\, d\Gamma + \right)\ \widetilde{x}_i\\
&\qquad + \int_\Omega \widetilde{V}_j\, \pi\, d\Omega\ + \int_\Gamma \widetilde{V}_j\, (\varphi + hT_f)\, d\Gamma
\end{aligned} \tag{7}$$

After the spatial discretization, the function $\widetilde{V}_i(M)$ becomes a vector $\widetilde{\mathbf{V}}_i\,[N,1]$ resulting in:

$$\forall\ j \in [1:n], \quad \sum_{i=1}^{n} \widetilde{\mathbf{V}}_j^{\ t}\ \mathbf{C}\ \widetilde{\mathbf{V}}_i\ \frac{d\widetilde{x}_i}{dt} = -\sum_{i=1}^{n} \widetilde{\mathbf{V}}_j^{\ t}\ \mathbf{A} \widetilde{\mathbf{V}}_i\ \widetilde{x}_i + \widetilde{\mathbf{V}}_j^{\ t}\ \mathbf{U} \tag{8}$$

We name $\widetilde{\mathbf{V}}[N,n]$ the matrix which gathers the n discretized functions $\widetilde{V}_i[N,1]$, and $\widetilde{\mathbf{X}}(t)[n,1]$ the vector of the n time-dependant excitation states $\widetilde{x}_i(t)$ associated with these space functions:

$$\widetilde{\mathbf{V}}^{\ t}\ \mathbf{C}\ \widetilde{\mathbf{V}}\ \frac{d\widetilde{\mathbf{X}}}{dt} = \widetilde{\mathbf{V}}^{\ t}\ \mathbf{A}\ \widetilde{\mathbf{V}}\ \widetilde{\mathbf{X}} + \widetilde{\mathbf{V}}^{\ t}\ \mathbf{U} \tag{9}$$

Under compact form:

$$\mathbf{L}\ \frac{d\widetilde{\mathbf{X}}}{dt} = \mathbf{M}\ \widetilde{\mathbf{X}} + \mathbf{N} \tag{10}$$

with $\mathbf{L} = \widetilde{\mathbf{V}}^{\ t}\ \mathbf{C}\ \widetilde{\mathbf{V}}$ and $\mathbf{M} = \widetilde{\mathbf{V}}^{\ t}\ \mathbf{A}\ \widetilde{\mathbf{V}}$ whose dimensions are $[n,n]$, and $\mathbf{N} = \widetilde{\mathbf{V}}^{\ t}\ \mathbf{U}\ [n,1]$.

This formulation leads to the reduction of the DOF, because the complete model (2) is characterized by N unknowns, while the dimension of this modal model (10) corresponds to the n space functions $\widetilde{V}_i(M)$.

From this formulation, different methods exist to reduce a model:

- The principle of the POD (*Proper Orthogonal Decomposition*) is the identification of the space functions $\widetilde{V}_i(M)$ from several reference temperature fields (noted $T_{ref}(M,t)$ for a thermal problem). This technique has been used in a lot of studies [5, 6, 7, 8, 9, 10, 11, 12].
- The MIM (*Modal Identification Method*) is based on the direct identification of the state equation providing with the modal formulation (10) from simulations or measures. This technique has been widely used for inverse problems [13, 14, 15, 16, 13, 17, 18].
- the PGD (*Proper Generalized decomposition*) is a generalization of the decomposition principle: the temperature is written as a multiple product of a set of functions, where each of these functions depends on one variable (time, space) or one parameter (heat capacity,thermal conductivity,...). These functions are computed in enriching the basis at each iteration [19, 20, 21, 22].

- The AROMM method follows both steps which appear in the modal principle, that is:
  - to compute a complete basis $\{V_i(M)\}_{i\in\mathbb{N}}$, on which it is possible to proceed to a rigorous decomposition of the thermal fields:

$$T(M,t) = \sum_{i=1}^{\infty} V_i(M)\, x_i(t) \tag{11}$$

  - to obtain a reduced basis $\{\tilde{V}_i(M)\}_{i\in[1,n]}$, in order to decrease the model order[3], and which allows to obtain a satisfactory estimation of the thermal field :

$$T(M,t) \simeq \sum_{i=1}^{n} \widetilde{V}_i(M)\, \widetilde{x}_i(t) \tag{12}$$

The goal of this lecture consists in presenting this method.

4 The complete basis computation

We search a set of spatial functions which form a basis for the considered thermal problem (11). This set depends on the solutions space.

4.1 Classical basis

4.1.1 The Fourier basis

We consider the following thermal problem, characterized by homogeneous boundary conditions:

$$\begin{cases} \forall M \in \Omega & : \quad c_0 \dfrac{\partial T}{\partial t} = \overrightarrow{\nabla}(k_0 \overrightarrow{\nabla} T) + \pi \\ \forall M \in \Gamma & : \quad k_0 \overrightarrow{\nabla} T.\overrightarrow{n} = -h_0\, T \end{cases} \tag{13}$$

The physical parameters (heat capacity c_0, thermal conductivity k_0, and global heat exchange coefficient h_0) are limited to spatial functions.
The space functions $\hat{V}_i^F(M)$ correspond to eigenvectors and are obtained by the resolution of the eigenvalues problem associated to the physical problem:

$$\begin{cases} \forall M \in \Omega & : \quad \overrightarrow{\nabla}\left(k_0 \overrightarrow{\nabla} \hat{V}_i^F\right) = z_i^F\, c_0\, \hat{V}_i^F \\ \forall M \in \Gamma & : \quad k_0 \overrightarrow{\nabla} \hat{V}_i.\overrightarrow{n} = -h_0\, \hat{V}_F \end{cases} \tag{14}$$

$z_i^F\ [\mathrm{s}^{-1}]$ is the eigenvalue associated to each eigenvector $\hat{V}_i^F$. The inverse of this quantity is a time $\tau_i^F\ [s]$ named the time constant of the eigenvector. It characterizes the dynamic of the eigenmode:

$$\tau_i^F = \frac{-1}{z_i^F} \tag{15}$$

These Fourier eigenmodes (Figure 1.a) can be considered as particular temperature fields: the eigenvalues problem corresponds to a stationary physical problem with a volumetric thermal load which is proportional to the eigenmode searched at each point of the domain, and whose boundary conditions are homogeneous.

[3] As we'll see later, the reduced function $\tilde{V}_i(M)$ do not correspond necessary with the functions $V_i(M)$ of the complete basis. This explains the change of notation

The variational form of the eigenvalues problem is :

$$\forall g \in H^1(\Omega), \quad -\int_\Omega k_0 \overrightarrow{\nabla} g \,.\, \overrightarrow{\nabla} \hat{V}_j^F \, \partial\Omega - \int_\Gamma g \, h_0 \, \hat{V}_i^F = z_i^F \int_\Omega g \, c_0 \, \hat{V}_i^F \, \partial\Omega \tag{16}$$

In cases of complex geometries, such eigenvalues problem is solved numerically, from a spatial discretization characterized by N DOF. The number of eigenmodes becomes then finite and equal to N. The numerical resolution is performed by the Lanczos method [23], from the discrete formulation of (16). In using the same matrix than specified previously (3), we have:

$$\mathbf{A} \; \hat{\mathbf{V}}_i^F = z_i^F \; \mathbf{C} \; \hat{\mathbf{V}}_i^F \tag{17}$$

This method had been implemented in all principal languages (Matlab since 1996 [24], Arpack since 1998 [25]). It allows to compute the eigenmodes according to the order of the most important time constants τ_i^F.

The set of all the eigenmodes $\hat{V}_i^F$ form a basis for the subspace $H_F^1(\Omega) \subset H^1(\Omega)$, which corresponds to this of the physical problem (13).

The eigenmodes are mutually-orthogonal according to a scalar product $< u, v >= \int_\Omega u c v \, \partial\Omega$:

$$\forall i \neq j, \quad < \hat{V}_i^F, \hat{V}_i^F >= \int_\Omega \hat{V}_i^F \, c_0 \, \hat{V}_j^F \, \partial\Omega = 0 \tag{18}$$

A standardization allows to impose the magnitude of each mode. In choosing:

$$V_i^F = \frac{\hat{V}_i^F}{\left(\int_\Omega \hat{V}_i^F \, c_0 \, \hat{V}_i^F \, d\Omega \right)^{1/2}} \tag{19}$$

we obtain the first orthogonality property:

$$\forall i, j \in \mathbb{N}, \quad < V_i^F, V_j^F >= \int_\Omega V_i^F \, c_0 \, V_j^F \, \partial\Omega = \delta_{ij} \tag{20}$$

Because of (16), and in choosing the eigenmodes V_j^F as test function, we have:

$$-\int_\Omega k_0 \overrightarrow{\nabla} V_i^F \,.\, \overrightarrow{\nabla} V_j^F \, d\Omega - \int_\Gamma V_i^F \, h_0 \, V_j^F \, d\Gamma = z_i^F \int_\Omega V_i^F \, c_0 \, V_j^F d\Omega \tag{21}$$

The use of the first orthogonality property (20) enables finally to obtain the second orthogonality property:

$$-\int_\Omega k_0 \overrightarrow{\nabla} V_i^F \,.\, \overrightarrow{\nabla} V_j^F \, d\Omega - \int_\Gamma V_i^F \, h_0 \, V_j^F \, d\Gamma = z_i^F \, \delta_{ij} \tag{22}$$

We saw previously that the state equation has been obtained by the projection of the thermal problem on the reduced basis (eq. (7)).
In the case where all the complete basis $\left(z_i^F \, , \, V_i^F \right)$ is used, we obtain:

$$\begin{aligned} \forall \, j \in \mathbb{N}, \\ \sum_{i=1}^{n} \left(\int_\Omega V_j^F \;\; c_0 \;\; V_i^F \, d\Omega \right) \;\; \frac{\partial x_i}{dt} = \\ + \sum_{i=1}^{n} \left(\int_\Omega k_0 \overrightarrow{\nabla} V_j^F . \overrightarrow{\nabla} V_i^F \, \partial\Omega \int_\Gamma V_j^F \, h_0 \, V_i^F \, \partial\Gamma + \right) \;\; x_i \\ + \int_\Omega \pi \, V_j^F \, \partial\Omega \end{aligned} \tag{23}$$

Because of the orthogonality properties (eq. (20) et (22)), all the state equations are fully decoupled:

$$\forall\, j \in \mathbb{N}, \quad \frac{\partial x_j}{\partial t} = z_j^F\, x_j + \int_\Omega V_j^F\, \pi\, \partial\Omega \tag{24}$$

As we will see later, the reduced basis $\left(\tilde{z}_i^F\ ,\ \tilde{V}_i^F\right)$ from his complete basis $\left(z_i^F\ ,\ V_i^F\right)$ is built, such as these previous orthogonality properties (eq. (20) et (22)) are preserved. The decoupled state-reduced equations (24) allow to obtain an immediate resolution.
The Fourier basis is valid for a linear thermal problem, with stationary parameters and with homogeneous boundary conditions, whatever the value of the thermal exchange coefficient $h_0(M)$.

In the particular case where $\forall M \in \Gamma, h_0 = 0$, we have the Neumann problem :

$$\begin{cases} \forall M \in \Omega & : \quad c_0 \dfrac{\partial T}{\partial t} = \overrightarrow{\nabla}(k_0 \overrightarrow{\nabla} T) + \pi \\ \forall M \in \Gamma & : \quad \overrightarrow{\nabla} T . \overrightarrow{n} = 0 \end{cases} \tag{25}$$

The eigenvalues problem associated is then the Neumann eigenvalues problem :

$$\begin{cases} \forall M \in \Omega & : \quad \overrightarrow{\nabla}\left(k_0 \overrightarrow{\nabla} \hat{V}_i^N\right) = z_i^N\, c_0\, \hat{V}_i^N \\ \forall M \in \Gamma & : \quad \overrightarrow{\nabla} \hat{V}_i^N\, .\, \overrightarrow{n} = 0 \end{cases} \tag{26}$$

This set of eigenvectors $\hat{V}_i^N$ forms a basis for the subspace $H_N^1(\Omega) \subset H^1(\Omega)$. Then they are characterized by a zero heat flux on the boundaries (Figure 1.b).

4.1.2 The Dirichlet basis

We consider a Dirichlet problem characterized by the following equations [4]:

$$\begin{cases} \forall M \in \Omega & : \quad c_0 \dfrac{\partial T}{\partial t} = \overrightarrow{\nabla}.(k_0 \overrightarrow{\nabla} T) + \pi \\ \forall M \in \Gamma & : \quad T = 0 \end{cases} \tag{27}$$

This problem defines a particular space of solutions named Dirichlet space H_0^1. It is a subspace of the Hilbert space H^1, which respects the boundary condition.

Eigenvectors $\hat{V}_i^D(M)$ are obtained by the resolution of the following eigenmodes problem :

$$\begin{cases} \forall M \in \Omega & : \quad \overrightarrow{\nabla}\left(k_0 \overrightarrow{\nabla} \hat{V}_i^D\right) = z_i^D\, c_0\, \hat{V}_i^D \\ \forall M \in \Gamma & : \quad \hat{V}_i^D = 0 \end{cases} \tag{28}$$

The variational form is as follows[5]:

[4] In practical terms, it is numerically possible to approach a Dirichlet thermal problem by a general Fourier formulation, in which we fix $h_0 \to \infty$. It is the same for the associated Dirichlet eigenvalues problem. Even if mathematical proof needs a rigorous writing of the problem (equations (27) and (28)), using such an expression for a numerical approach gives good results.

[5] One use here a test function $g \in H_0^1(\Omega)$, which has then a zero value on the boundaries. The integral term $\int_\Gamma g\, k\, \overrightarrow{\nabla} \hat{V}_i^D . \overrightarrow{n}\, d\Gamma$ is then a zero value.

$$\forall g \in H_0^1(\Omega), \quad -\int_\Omega k_0 \overrightarrow{\nabla} g \,.\, \overrightarrow{\nabla} \hat{V}_j^D \, \partial\Omega = z_i \int_\Omega g \, c_0 \, \hat{V}_i^D \, \partial\Omega \tag{29}$$

This set of all the eigenvectors V_i^D forms a basis for the Dirichlet subspace $H_0^1(\Omega) \subset H^1(\Omega)$. Then they are characterized by a zero value on the boundaries, as shown in figure (1.c).

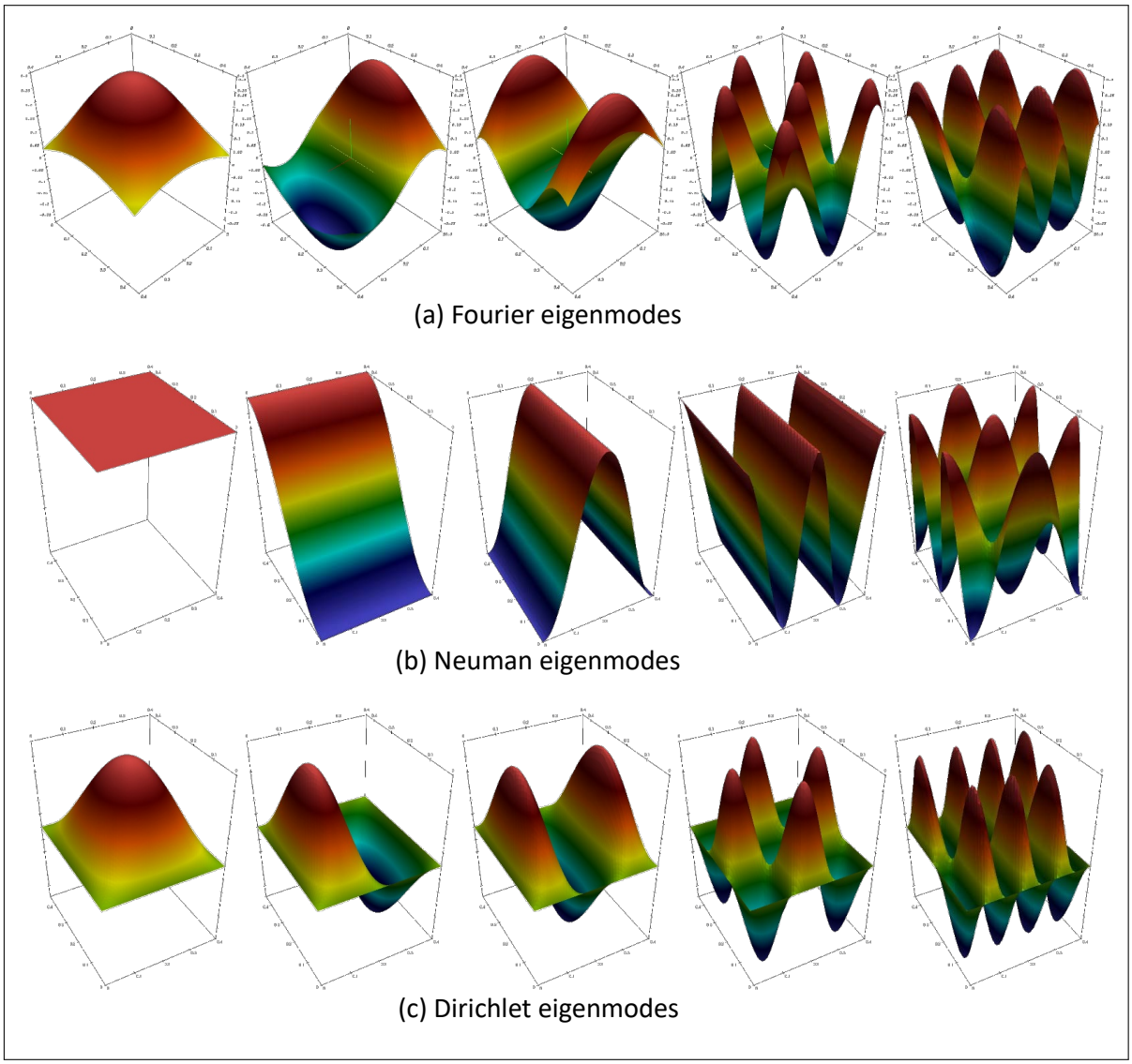


Figure 1: Classical modes for a simple 2D rectangular geometry

An adapted normalization[6] enables to fix the magnitude of the modes, and leads to the following orthogonality relations:

$$\forall i,j \in \mathbb{N}, \begin{cases} \int_\Omega V_i^D \, c_0 \, V_j^D \, \partial\Omega = \delta_{ij} \\ \int_\Omega k_0 \overrightarrow{\nabla} V_i^D \, . \, \overrightarrow{\nabla} V_j^D \, d\Omega = z_i^D \, \delta_{ij} \end{cases} \tag{30}$$

4.1.3 Non homogeneous problem: applying a gliding temperature

We consider the general problem for which we recall the equations:

$$\begin{cases} \forall M \in \Omega & : & c_0 \dfrac{\partial T}{\partial t} = \overrightarrow{\nabla}(k_0 \overrightarrow{\nabla} T) + \pi \\ \forall M \in \Gamma & : & k_0 \overrightarrow{\nabla} T . \overrightarrow{n} = \varphi + h_0 (T_f - T) \end{cases} \tag{31}$$

We saw that the Fourier eigenmodes (14) form a basis for a thermal problem characterized by homogeneous boundary conditions. In order to use the modal reduction with theses eigenmodes, we have to split the temperature T on two terms :

$$T = T_g + T_d \tag{32}$$

- The term T_g is called the gliding temperature, because it corresponds to the temperature obtained without any consideration of the thermal inertia:

$$\begin{cases} \forall M \in \Omega & : & 0 = \overrightarrow{\nabla}(k_0 \overrightarrow{\nabla} T_g) + \pi \\ \forall M \in \Gamma & : & k_0 \overrightarrow{\nabla} T_g . \overrightarrow{n} = \varphi + h_0 (T_f - T_g) \end{cases} \tag{33}$$

  Such a problem is simple: from the variationnal formulation from (33):

$$-\int_\Omega k_0 \overrightarrow{\nabla} g \, . \, \overrightarrow{\nabla} T_g \, d\Omega - \int_\Gamma g \, h_0 \, T_g \, d\Gamma + \int_\Omega g \, \pi \, d\Omega + \int_\Gamma g \, (\varphi + h_0 \, T_f) \, d\Gamma = 0 \tag{34}$$

  the discrete form is then:

$$\mathbf{A}\mathbf{T_g} + \mathbf{U}(t) = 0 \tag{35}$$

  and we have then:

$$\mathbf{T_g} = -\mathbf{A}^{-1}\mathbf{U}(t) \tag{36}$$

- The complementary variable T_d is called the dynamic temperature. From (31) and (33), the equation which allows to obtain T_d is :

$$\begin{cases} \forall M \in \Omega & : & c_0 \dfrac{\partial T_d}{\partial t} = \overrightarrow{\nabla}(k_0 \overrightarrow{\nabla} T_d) - c_0 \dfrac{\partial T_g}{\partial t} \\ \forall M \in \Gamma & : & k_0 \overrightarrow{\nabla} T_d . \overrightarrow{n} = -h_0 \, T_d \end{cases} \tag{37}$$

  Such problem is then homogeneous and it is then allowed to reduce it by using the Fourier basis.

Lastly the researched temperature field T is:

$$T = \sum_{i=1}^{\infty} x_i \, V_i^F + T_g \tag{38}$$

[6] It is the same as the one used for the Fourier eigenmodes (eq. (19))

The state modal problem is always decoupled. The gliding temperature T_g appears only in cases of time variation of the solicitations:

$$\forall(i) \in \mathbb{N}, \quad \frac{dx_i}{dt} = z_i\, x_i - \int_\Omega V_i^F\, c_0 \frac{dT_g}{dt} d\Omega \tag{39}$$

Several studies have used this technique, including buildings problems [26, 27, 28, 29].

Hovewer, the limit of this method is that the computed basis is applicable only for problems in which the boundary conditions are fixed. From the second equation of (14), we can define the quantity γ_i such as:

$$\gamma_i = \frac{\overrightarrow{\nabla} V_i . \overrightarrow{n}}{V_i} = \frac{-h_0}{k_0} \tag{40}$$

In this way we can see that all the eigenvectors are characterized by the same value of this quantity γ_i. Thus, all the dynamic thermal fields that can be rebuilt by this modal formulation have to respect this constraint.
Such basis are not compatible with a thermal problem in which non linearities or time variations exist on the boundaries. Examples are numerous: time dependant exchange coefficient $h(t)$, thermal conductivity depending on the temperature $k(T)$, infrared radiations... That is why other basis have been developed.

4.2 Basis adapted to non linear problems

4.2.1 Branch modes

In order to avoid this limit, a new basis is defined, whose boundary conditions are not linked with the physical boundary conditions:

$$\begin{cases} \forall M \in \Omega \quad , \quad k_0 \overrightarrow{\nabla}(\overrightarrow{\nabla} \hat{V}_i^B) = z_i^B\, c_0\, \hat{V}_i^B \\ \forall M \in \Gamma \quad , \quad k_0 \overrightarrow{\nabla} \hat{V}_i^B . \overrightarrow{n} = -z_i^B\, \zeta\, \hat{V}_i^B \end{cases} \tag{41}$$

The feature of this basis is that the eigenvalues z_i^B is present in the boundary condition. This is the Steklov condition.
The quantity ζ [J.m^{-2}K^{-1}] is called Steklov parameter and it is a simple coefficient which allows to respect the physical dimensions in the boundary condition equations. The value of this coefficient is obtained from the variational formulation of the eigenvalues problem (41).

$$-\int_\Omega k_0 \overrightarrow{\nabla} g . \overrightarrow{\nabla} V_i^B\, d\Omega = z_i \left(\int_\Omega c_0\, g\, V_i^B\, d\Omega + \int_\Gamma \zeta\, g\, V_i^B\, d\Gamma \right) \tag{42}$$

To balance the two terms linked to the eigenvalue, an appropriate choice of the Steklov coefficient ζ is given by:

$$\zeta \simeq \frac{\int_\Omega c_0\, d\Omega}{\int_\Gamma d\Gamma} \tag{43}$$

In using the associated scalar product:

$$< u, v >= \int_\Omega u\, c_0\, v\, d\Omega + \int_\Gamma u\, \zeta\, v\, d\Gamma \tag{44}$$

the normalisation is done:

$$V_i^B = \frac{\hat{V}_i^B}{\left(\int_\Omega \hat{V}_i^B \, c_0 \, \hat{V}_i^B \, d\Omega + \int_\Gamma \hat{V}_i^B \, \zeta \, \hat{V}_i^B \, d\Gamma\right)^{1/2}} \tag{45}$$

and we obtain the following orthogonality properties:

$$\begin{aligned} &\forall (i,j) \in \mathbb{N}^2, \\ &\int_\Omega V_j^B \, c_0 \, V_i^B \, d\Omega + \int_\Gamma V_i^B \, \zeta \, V_i^B \, d\Gamma = \delta_{ij} \\ &\int_\Omega k_0 \overrightarrow{\nabla} V_j^B \, . \, \overrightarrow{\nabla} \, V_i^B \, d\Omega = z_i^B \, \delta_{ij} \end{aligned} \tag{46}$$

It is possible to characterize the spatial evolution of each Branch modes by defining a form coefficient C_i^ζ for each mode V_i^B:

$$C_i^\zeta = \int_\Gamma V_i^B \, \zeta \, V_i^B \, d\Gamma \tag{47}$$

The evolution of this coefficient according to the mode number, for a simple rectangular geometry, is presented on figure 2. It shows that two Branch modes families exist:

- Because of the orthogonality relation defined in Eq. (14), when C_i^ζ is close to 1, the considered mode is flat on the domain except near the border. Such modes are called Boundary modes. They do not appear in a classical Fourier basis, and allow the reconstitution of any boundary conditions.
- There exist others modes for which the spatial evolutions are located in all the domain. We call them Domain modes. These modes are characterized by a weak value of c_i^ζ (less than 0.3 for the example in figure 2). These are less numerous as the Boundaries modes (for the first computed modes).

Figure 3.a represents some Branch modes for a simple 2D rectangular geometry. This figure enables to clearly visualize these two families of Branch modes.

With these Branch modes, the orthogonality properties don't allow anymore to obtain a decoupled modal problem:

$$\begin{aligned} \forall j \in \mathbb{N}, \quad &\sum_{i=1}^{\infty} \left(\int_\Omega V_j^B \, c \, V_i^B \, d\Omega \right) \frac{dx_i}{dt} \\ &= \sum_{i=1}^{\infty} \left(\int_\Omega k \overrightarrow{\nabla} V_j^B \, . \overrightarrow{\nabla} V_i^B \, d\Omega + \int_\Gamma V_j^B \, h \, V_i^B \, d\Gamma \right) x_i \\ &+ \int_\Omega V_j^B \, \pi \, d\Omega + \int_\Gamma V_j^B \, (hT_e + \varphi) d\Gamma \end{aligned} \tag{48}$$

This is the price to pay for using this Branch basis.

On the other hand, the Branch modes form a basis for any thermal problem, including those characterized by parameters that are functions of time or temperature. One shows that the generated functionnal space is the Hilbert space $H^1(\Omega)$ and we have directly[7]:

[7] It is no longer necessary to use the sliding temperature field

$$T(M,t) = \sum_{i=1}^{\infty} x_i \, V_i^B \tag{49}$$

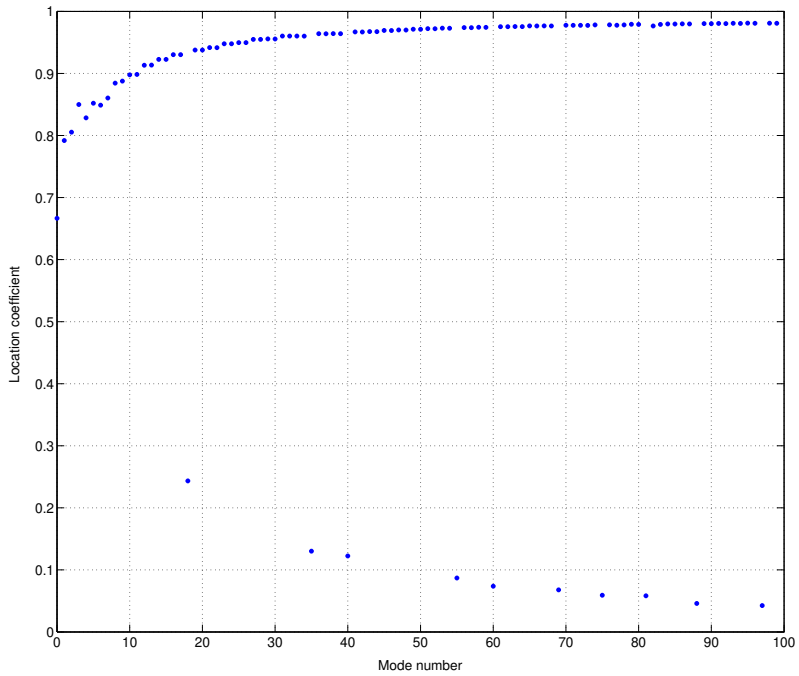


Figure 2: Evolution of the location coefficient according to the branch mode number

Initiated by Neveu *et al.* [30], this base type has been applied to different configurations: Quéméner *et al.* [31] treats the case of a non-linear problem, with the existence of solidification of a molded part. Various applications are made for inverse problems by Videcoq *et al.* [32, 33, 34]. Branch bases generalized to problems of diffusion with transport are proposed by Joly *et al.* [35], then used in the case of an inverse problem of identification [36]. Finally Laffay *et al.* [37, 38] proposes a substructuring technique, which allows the computation of Branch bases for different subdomains, which are then coupled each other by a thermal contact resistance.

4.2.2 The Dirichlet-Steklov eigenmodes

Recently another way to reduce non linear problems with or without time dependant parameters has been developped. It consist in using two bases:

- the Dirichlet basis seen previously (eq. (27)),
- the Steklov basis[8], which is defined by the following eigenvalues problem:

$$\begin{cases} \forall M \in \Omega \quad , \quad \overrightarrow{\nabla}(k_0 \overrightarrow{\nabla} \hat{V}_i^S) = 0 \\ \forall M \in \Gamma \quad , \quad k_0 \overrightarrow{\nabla} \hat{V}_i^S . \overrightarrow{n} = -z_i^S \, \zeta \, \hat{V}_i^S \end{cases} \tag{50}$$

Steklov modes correspond to stationary fields obtained for a problem in which one imposes fluxes at the boundaries, whose value is proportional to the value of this mode at

[8] Steklov modes are rigorously defined only on the boundaries. In order to simply the notation, we call here by abuse of language the steklov mode as their extension in the domain (noted $\hat{V}_i^S$)

any point on the border.

The regrouping of these two families of modes $\{V_i^D\}_{i\in\mathbb{N}} \bigoplus \{V_j^S\}_{j\in\mathbb{N}}$ forms a hilberian basis de $H^1(\Omega)$.
We define the following scalar product:

$$< u, v >= \int_\Omega k_0 \overrightarrow{\nabla} u \ . \ \overrightarrow{\nabla} v \, d\Omega + z_0 \int_\Gamma u \, \zeta \, v \, d\Gamma \tag{51}$$

where z_0 is a constant parameter $[s^{-1}]$ which allows to respect the coherence of the physical dimension of both terms.

Using the following standardization:

$$V_i^S = \frac{\hat{V}_i^{DS}}{\left(\int_\Omega k_0 \overrightarrow{\nabla} \hat{V}_i^{DS} \, . \, \overrightarrow{\nabla} \hat{V}_i^{DS} \, d\Omega + z_0 \int_\Gamma \hat{V}_i^{DS} \, \zeta \, \hat{V}_i^{DS} \, d\Gamma \right)^{1/2}} \tag{52}$$

we obtain Dirichlet and Steklov modes which are orthogonal with respect to this scalar product (51):

$$\begin{aligned} &\forall \mathcal{X}, \mathcal{Y} \in \{D, S\}, \forall i, j \in \mathbb{N}, \\ &< \hat{V}_i^{\mathcal{X}}, \hat{V}_j^{\mathcal{Y}} > \ = \int_\Omega k_0 \overrightarrow{\nabla} \hat{V}_i^{\mathcal{X}} \, . \, \overrightarrow{\nabla} \hat{V}_j^{\mathcal{Y}} \, d\Omega + z_0 \int_\Gamma \hat{V}_i^{\mathcal{X}} \, \zeta \, \hat{V}_j^{\mathcal{Y}} \, d\Gamma \\ &\qquad\qquad = \delta_{\mathcal{X}\mathcal{Y}} \, \delta_{ij} \end{aligned} \tag{53}$$

A sets of modes of the Dirichlet-Steklov basis is compared to the Branch modes in Figure 3. This shows that Steklov's modes correspond very well to Boudaries Branch modes, whereas Domain Branches modes and Dirichlet modes are similar only inside the domain. At the boundaries,the Domain Branch modes are not characterized by null values, unlike Dirichlet modes. Nevertheless the correspondence between these two bases is flagrant.

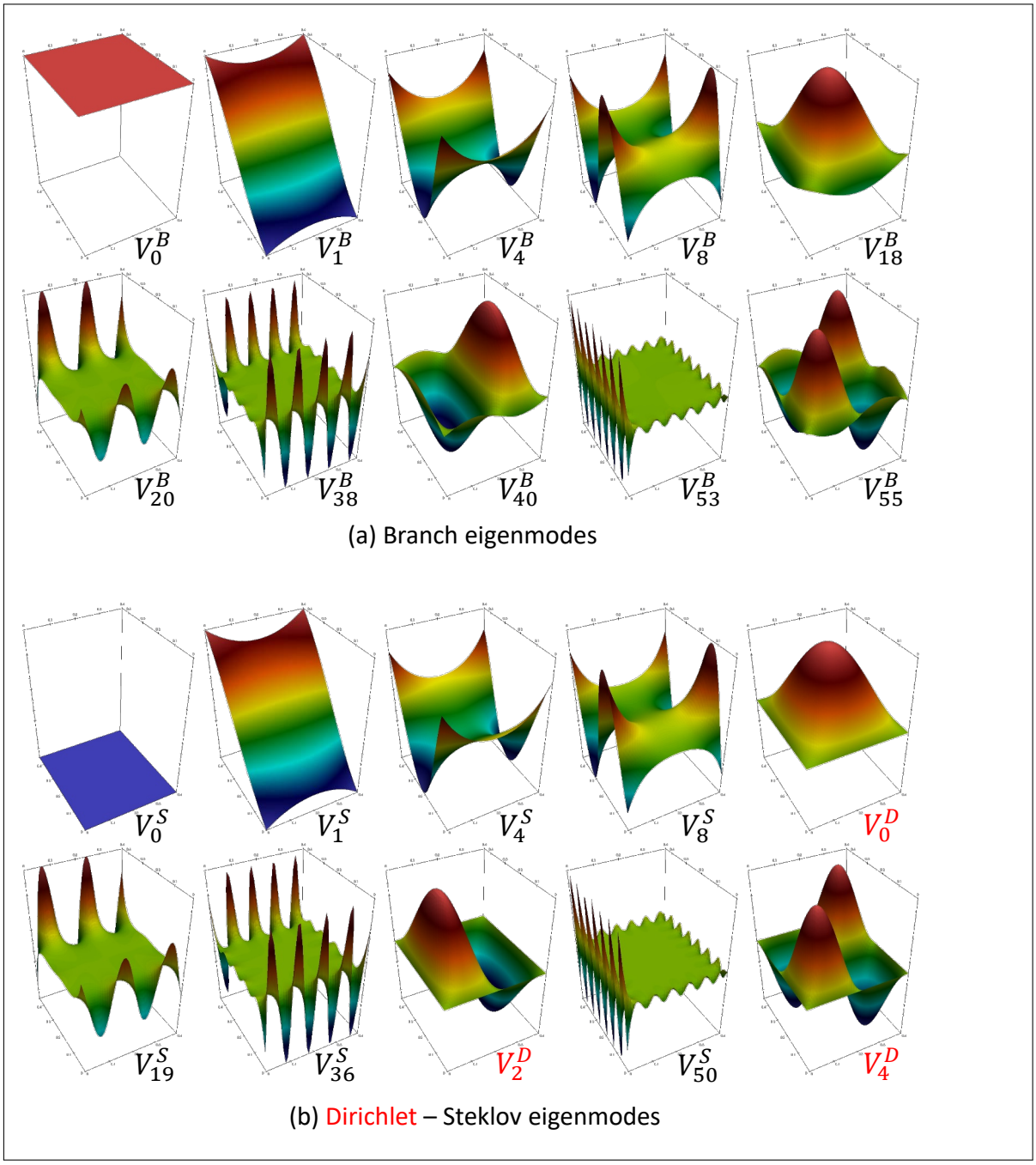


Figure 3: Comparison between the Branch basis $\{V_i^B\}$ and the Dirichlet-Steklov basis $\{V_i^D\} \bigoplus \{V_j^S\}$

5 Reducing the basis

Until now, no reduction has been made. Whatever the chosen basis, the problem of state (eq.(39) or (48)) remains characterized by a size related to spatial discretization. The second step of the AROMM method is then to build a reduced base containing n modes $\tilde{V}_i(M)$ from the complete base. We saw previously that the form of the modal problem resulting from this reduction depends on the used base:

- For a base associated with a linear thermal problem and with stationary parameters (ie Fourier base $\{V_i^F\}$, Neumann base $\{V_i^N\}$ or Dirichlet base $\{V_i^D\}$):

$$\forall i \in \{1,n\} \quad \frac{dx_i}{dt} = z_i\, x_i - \int_\Omega \tilde{V}_i\, c_0\, \frac{dT_g}{dt} d\Omega \tag{54}$$

- For a base adapted to more general problems (ie Branch base $\{V_i^B\}$ or Diriclet-Steklov base $\{V_i^D\} \bigoplus \{V_j^S\}$):

$$\begin{aligned} \forall j \in \{1,n\} \quad & \sum_{i=1}^{n} \left(\int_\Omega \tilde{V}_j\, c\, \tilde{V}_i\, d\Omega \right) \dot{x}_i \\ & = \sum_{i=1}^{n} \left(\int_\Omega k \vec{\nabla} \tilde{V}_j \,.\, \vec{\nabla} \tilde{V}_i\, d\Omega + \int_\Gamma \tilde{V}_j\, h\, \tilde{V}_i\, d\Gamma \right) x_i \\ & + \int_\Omega \tilde{V}_j\, \pi\, d\Omega + \int_\Gamma \tilde{V}_j\, (h\, T_e + \varphi)\, d\Gamma \end{aligned} \tag{55}$$

Several reduction methods exist.

5.1 Truncation

The simplest idea is to take the most relevant modes from the complete base:

$$\forall i \in \{1,n\} \quad \forall j \in \{1,N\} \quad , \quad \tilde{V}_i = V_j \tag{56}$$

5.1.1 Temporal Truncation

A first criterion leads to the truncation of Marshall [39]. In this method the modes with the largest time constants are kept. Independent of any reference problem, this reduction technique has mostly been used for classical basis [40].

An important advantage of this reduction is that it is immediate to use, since the Lanczos technique allows the base to be calculated according to the order of the largest time constants. Thus, temporal truncation can also be used as first-level reduction: instead of calculating the complete base, only a certain percentage of this base is computed, from which it is possible to make a second reduction more efficient. In the case of thermal problems characterized by a very large number of DOF, this possibility of partial calculations of the base is of great interest, given the important calculation times needed for solving the eigenvalue problem and the difficulties of the eigenvectors storage.

5.1.2 Energetic Truncation

This technique is used by Joly *et al.* [35]. From a set of known temperature fields $T_{ref}(t)$, it is possible to obtain the excitation states by a simple projection of the complete basis on

T_{ref} according to the definition of the scalar product defined for the considered basis.

For example, in the case of Branch basis, orthogonal properties lead to:

$$\begin{aligned}
&\forall j \in \mathbb{N}, \\
&\int_\Omega T_{ref}\, c_0\, V_j^B\, d\Omega + \int_\Gamma T_{ref}\, \zeta\, V_j^B\, d\Gamma \\
&\qquad = \int_\Omega \sum_{i=1}^{n} \left(x_i\, V_i^B\right) c_0\, V_j^B\, d\Omega + \int_\Gamma \sum_{i=1}^{n} \left(x_i\, V_i^B\right)\, \zeta\, V_j^B\, d\Gamma \\
&\qquad = \sum_{i=1}^{n} \left(\int_\Omega V_i^B\, c_0\, V_j^B\, d\Omega + \int_\Gamma V_i^B\, \zeta\, V_j^B\, d\Gamma \right) x_i \\
&\qquad = \sum_{i=1}^{n} \delta_{ij} x_i \\
&\qquad = x_j
\end{aligned} \tag{57}$$

For the Dirichlet Steklov basis, given the definition of the scalar product used, the projection leads to:

$$\begin{aligned}
&\forall \mathcal{X}, \mathcal{Y} \in \{D, S\}, \forall j \in \mathbb{N}, \\
&\int_\Omega k_0 \overrightarrow{\nabla} T_{ref} \,.\, \overrightarrow{\nabla} \hat{V}_j^{\mathcal{Y}}\, d\Omega + \int_\Gamma T_{ref}\, \zeta\, \hat{V}_j^{\mathcal{Y}}\, d\Gamma \\
&\qquad = \int_\Omega k_0 \overrightarrow{\nabla} \left(\sum_{i=1}^{n} x_i\, \hat{V}_i^{\mathcal{X}} \right) \,.\, \overrightarrow{\nabla} \hat{V}_j^{\mathcal{Y}}\, d\Omega + \int_\Gamma \left(\sum_{i=1}^{n} x_i\, \hat{V}_i^{\mathcal{X}} \right)\, \zeta\, \hat{V}_j^{\mathcal{Y}}\, d\Gamma \\
&\qquad = \sum_{i=1}^{n} \left(\int_\Omega k_0 \overrightarrow{\nabla} \hat{V}_i^{\mathcal{X}} \,.\, \overrightarrow{\nabla} \hat{V}_j^{\mathcal{Y}}\, d\Omega + \int_\Gamma \hat{V}_i^{\mathcal{X}}\, \zeta\, \hat{V}_j^{\mathcal{Y}}\, d\Gamma \right) x_i \\
&\qquad = \sum_{i=1}^{n} \delta_{\mathcal{X}\mathcal{Y}}\, \delta_{ij}\, x_i \\
&\qquad = x_j
\end{aligned} \tag{58}$$

The knowledge of the excitation states for all the modes of the complete basis makes it possible to keep only those characterized by the most important states for all the temperature fields used. This technique generally leads to a more efficient reduction than the simple temporal truncation, but it has a disadvantage: the effectiveness of the reduction depends on the reference fields that must be known. Here we find the same constraint as that existing for the POD method.
From the same discretized geometry it is generally possible to perform simulations of a thermal problem which is simpler than that studied, but which will however be able to excite the characteristic modes.

5.2 Amalgamated base

An even more elaborate technique is that of the amalgam. It brings back the idea of classifying the eigenmodes according to their states of excitation, but this time, the modes which are not kept during the truncation are added by simple linear combinations to the retained modes:

$$\forall i \in \{1, n\} \quad \tilde{V}_i = V_{i,1} + \sum_{p=2}^{\tilde{N}_i} \alpha_{i,p} V_{i,p} \quad ; \quad 0 < |\alpha_{i,p}| < 1 \tag{59}$$

In order to maintain the properties of the base, each mode is used only once:

$$\sum_{i=1}^{n} \left(\tilde{N}_i + 1 \right) = N \tag{60}$$

The distribution of the initial modes and the computation of the amalgam coefficient $\alpha_{i,p}$ are carried out in a fast sequential procedure which depends only on the knowledge of the excitation states. Set up by Oulefki [41] in the case of classical bases for which decoupling made it easy to determine the reference states, this reduction technique has been widely used for Branch modes.
The difficulty is in general to determine the excitation states of the complete basis. A first rather simple solution [42, 43] is, as for energetic truncation, to use a set of temperature fields obtained by complete resolution of a reference problem, which gives access to the excitation states (eq. (57) or (58)).
Other techniques have also been tested [31, 33, 37] in order to avoid computing the reference thermal fields: since the eigenmodes excitation states are known only to classify this modes in order to set up the amalgam procedure, these authors have built the associated complete modal problem, and sought a simple estimate of the states of excitation: Using a Branch basis and neglecting the terms of coupling between modes, the modal problem has been solved analytically and the excitation states became extremely fast to obtain [31]. An improvement of this technique has been carried out later in the case of a rotating disc, for which only the coupling of a small number of modes is taken into account [44].

6 Application to the inverse problems: Examples

The examples presented here concern the automobile brake system, which is a major safety component. It undergoes, during its operating phase, many mechanical and thermal stresses, which can lead to important damages: cracks, apparition of hot-judder, vapor locking, brake fade, etc.
Because of thermal solicitations are rarely known (especially the part of the heat flux received by the pad and by the disc), the inverse techniques is used. In order to respect the complex geometry of the system, the model used in the inverse process is numerical, and characterized by very fine meshes. Computing time and memory problems appear very quickly, and a solution is to use reduced models.

6.1 Estimation of heat flux received by the brake disc rotating [1]

A brake disc in rotation with variable rotation frequency $\omega(t)$ is considered (Fig. 4). During the braking phase, the disc receives a time-dependent heat flux on the zone of friction with the brake pads Ω_1. The flux density $\varphi[W.m^{-2}]$ dissipated by friction is not uniform but varies linearly with the velocity thus with the radius.
The space discretization using P1 finite elements leads to a DOF $N = 9860$ for the following matrix formulation:

$$\mathbf{C}\frac{d\mathbf{T}}{dt} = \left[\mathbf{K} + \omega_u(t)\mathbf{U} + h_u(t)\mathbf{H}\right]\mathbf{T} + \varphi_u\mathbf{U} \tag{61}$$

The goal consists in identifying $\varphi_u(t)$ in real time, from a local infrared mesurement on the disc (point A).
Concerning the direct simulation, the computing time is significant (equal to 2160 s on a simple laptop), because of the transport term which involves small computation time-steps. Figure 5 illustrates this phenomenon.
Such simulation time is an obstacle for inverse applications where the need for real-time response is important. To avoid prohibitive time, a reduced model is built. It is obtained by

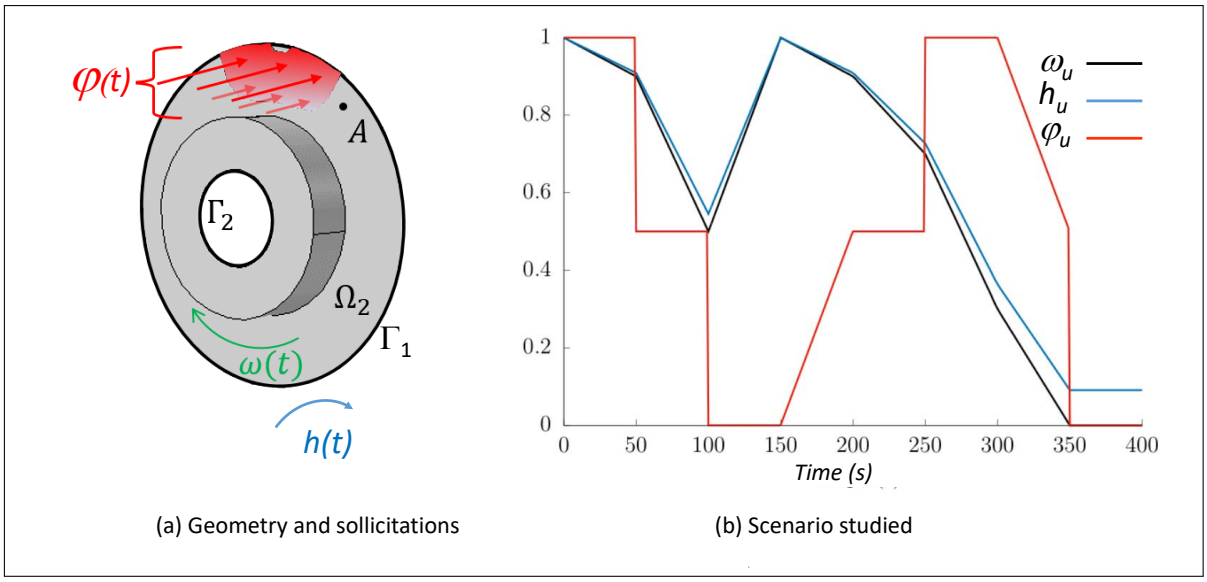


Figure 4: Physical problem

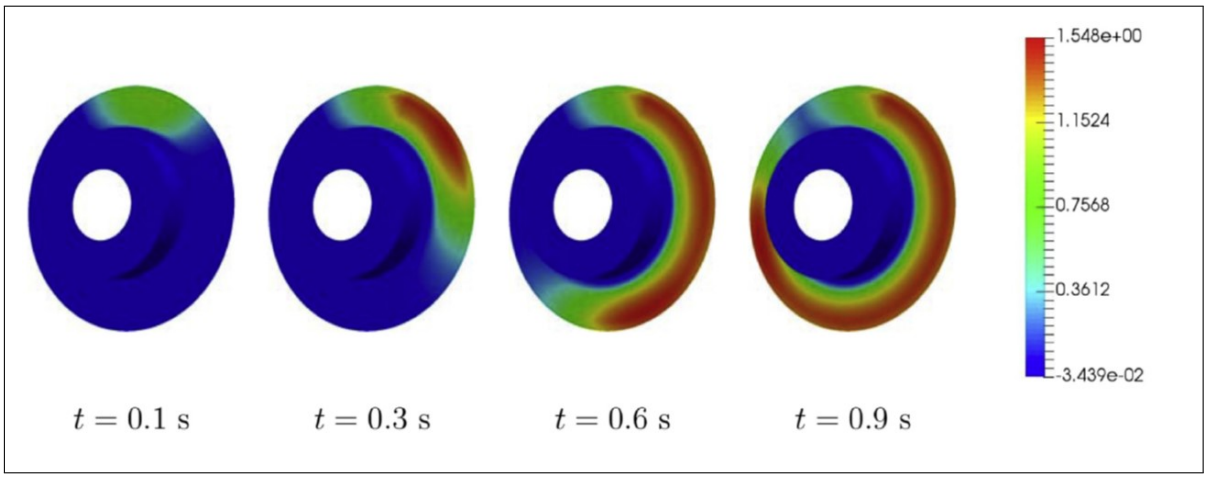
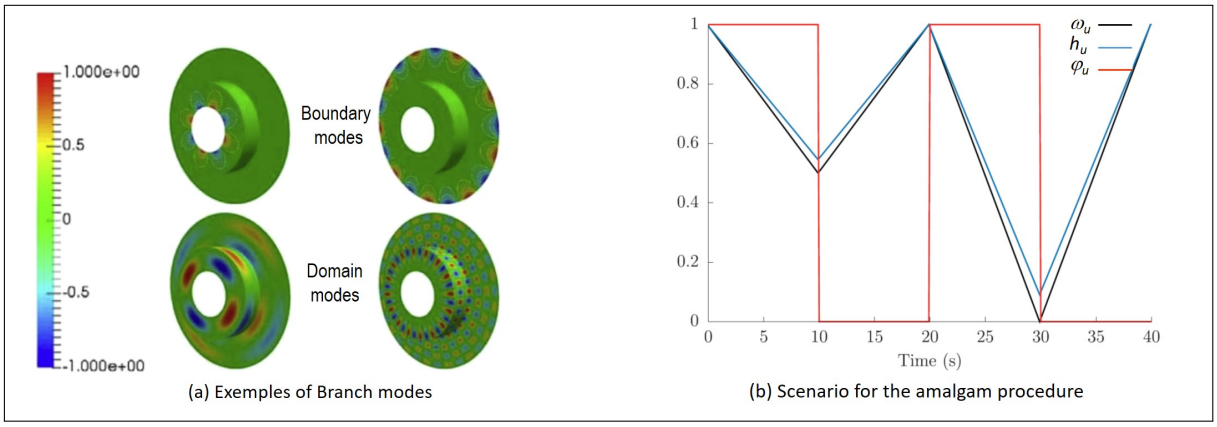


Figure 5: Temperature fields at different times

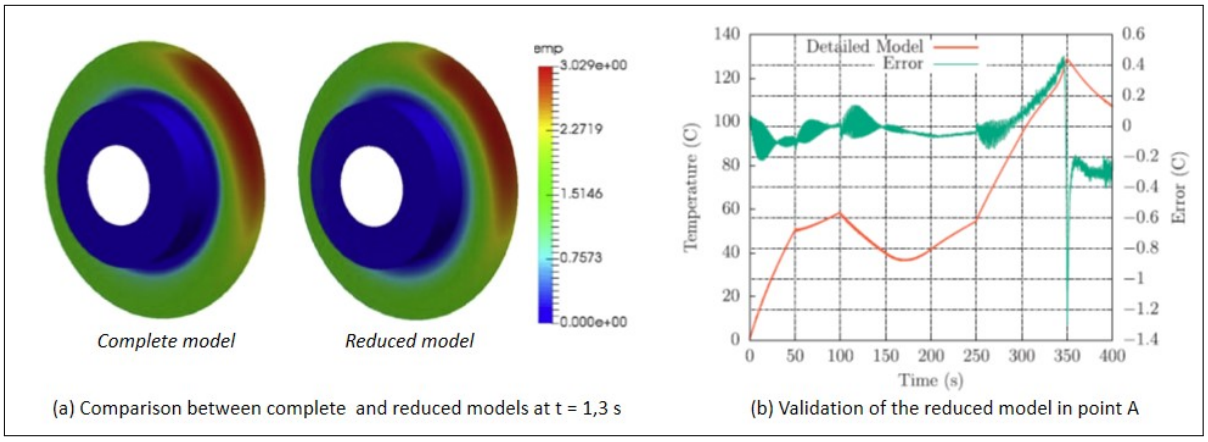
the AROMM method, in choosing Branch eigenmodes (fig 6.a) and the Amalgam method. The reference scenario used for the Amalgam procedure (fig 6.b) is obviously different from the one used for the identification (fig 4.b). With a reduced order $n = 15$, the direct simulation requires less than 10s, with satisfying results (fig 7).
By integrating such reduced model in an inverse approach, it is then possible to identify the heat flux φ in quasi real time. The inverse algorithm is based on the adjoint method applied on sliding time windows (fig 8).



(a) Exemples of Branch modes

(b) Scenario for the amalgam procedure

Figure 6: Reduced model



(a) Comparison between complete and reduced models at t = 1,3 s

(b) Validation of the reduced model in point A

Figure 7: Using the reduced model $\tilde{n} = 15$ in direct simulation

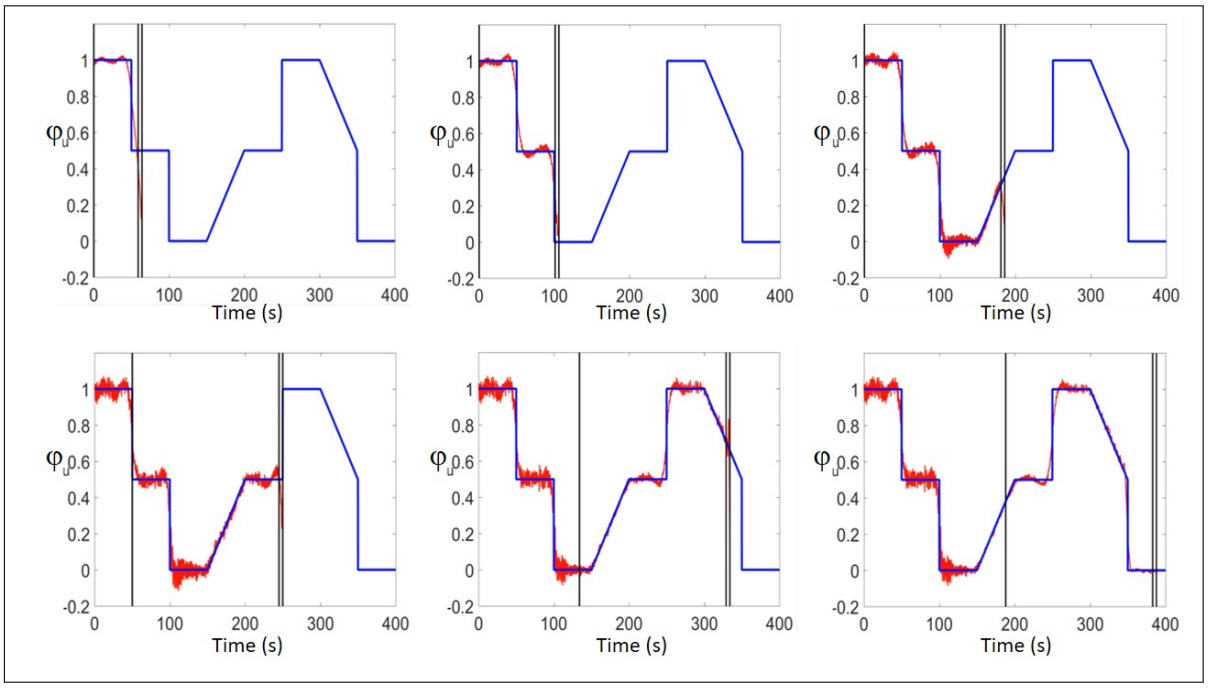


Figure 8: Identification results

6.2 Spatio-temporal identification of heat flux density received by the brake pad [2]

The identification of the spatio-temporal variations of a heat flux density field is addressed in this section. The application relates to the identification of the heat flux received by a brake pad in a braking situation, for which the mechanical deformation and the phenomena of tear and wear cause the appearance of hot spots that one seeks to locate.
We consider a car brake pad for which the complexity of the geometry is respected (Fig. 9.a). It is composed of two materials: the brake lining and its metallic support. This brake pad undergoes three types of boundary conditions (fig 9.b).

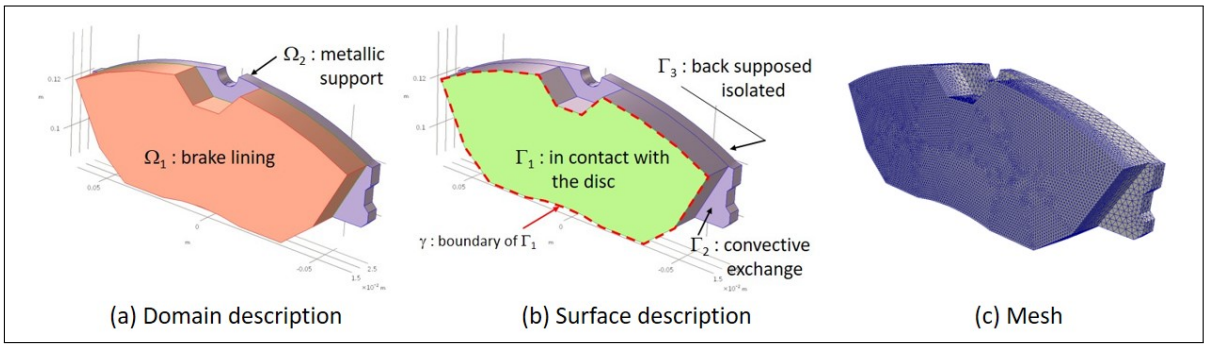


Figure 9: Geometry of the pad and its discretization

6.2.1 Parametrization of the heat flux density

A first Branch base $V^{(\varphi)}$ is used in order to parametrize the heat flux density (fig. 10):

$$\varphi(x,y,t) = \sum_{k=1}^{n^{(\varphi)}} \tilde{x}_k^{(\varphi)}(t)\, \tilde{V}_k^{(\varphi)}(x,y) \tag{62}$$

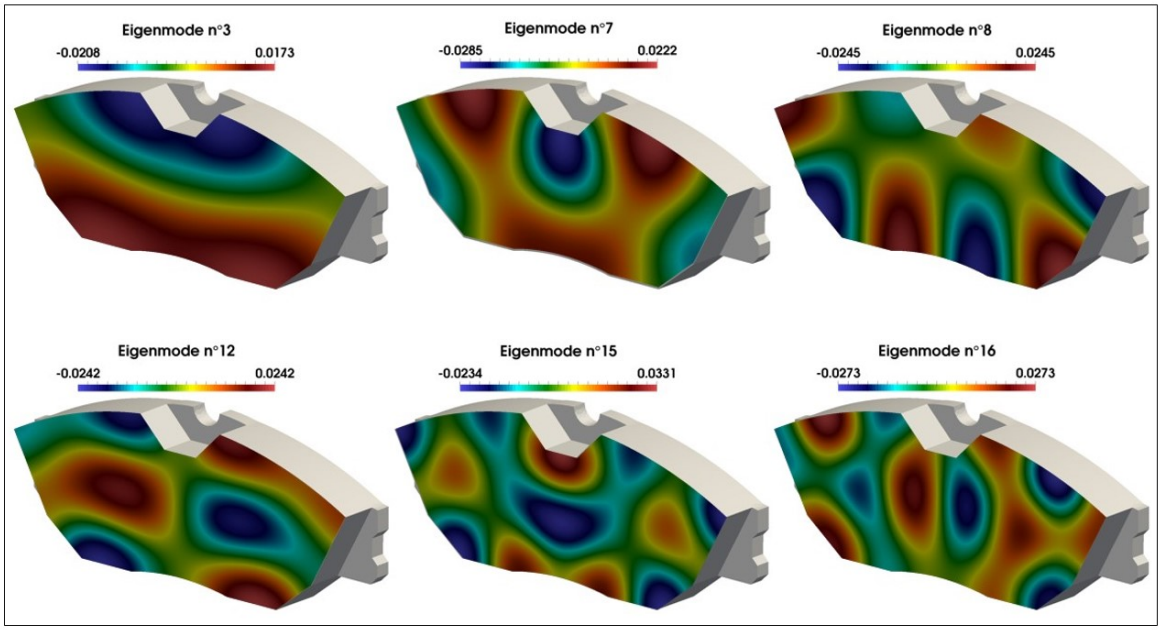


Figure 10: Flux basis

The equation of the heat discretized (eq. 3) becomes then:

$$\mathbf{C}\dot{\mathbf{T}} = (\mathbf{K}+\mathbf{H})\mathbf{T} + \sum_{k=1}^{n^{(\varphi)}} \mathbf{W}\,\tilde{\mathbf{V}}_k^{(\boldsymbol{\varphi})}\,\tilde{x}_k^{(\varphi)} \tag{63}$$

where $\tilde{\mathbf{V}}_k^{(\boldsymbol{\varphi})}$ $[N_{mesh} \times 1]$ is the extension on the domain Ω, of each eigenvector $\tilde{V}_k^{(\varphi)}$ computed on the boundary Γ_1, and where the matrix $\mathbf{W}$ $[N_{mesh} \times N_{mesh}]$ corresponds to the integration of the interpolations functions defined on the border Γ_1 and extended to the domain Ω.
This can be written compactly:

$$\mathbf{C}\dot{\mathbf{T}} = (\mathbf{K}+\mathbf{H})\mathbf{T} + \mathbf{W}\,\tilde{\mathbf{V}}^{(\boldsymbol{\varphi})}\,\tilde{\mathbf{X}}^{(\boldsymbol{\varphi})} \tag{64}$$

where $\tilde{\mathbf{V}}^{(\boldsymbol{\varphi})}$ is a matrix of dimension $[N_{mesh} \times n^{(\varphi)}]$ which gathers all the flux modes $\tilde{\mathbf{V}}_k^{(\boldsymbol{\varphi})}$ $[N_{mesh} \times 1]$ used, and $\tilde{\mathbf{X}}^{(\boldsymbol{\varphi})}$ is the vector of the corresponding states of dimension $[n^{(\varphi)} \times 1]$.

6.2.2 Reduced problem

A second Branch base V^T is used for the temperature field (fig. 11)

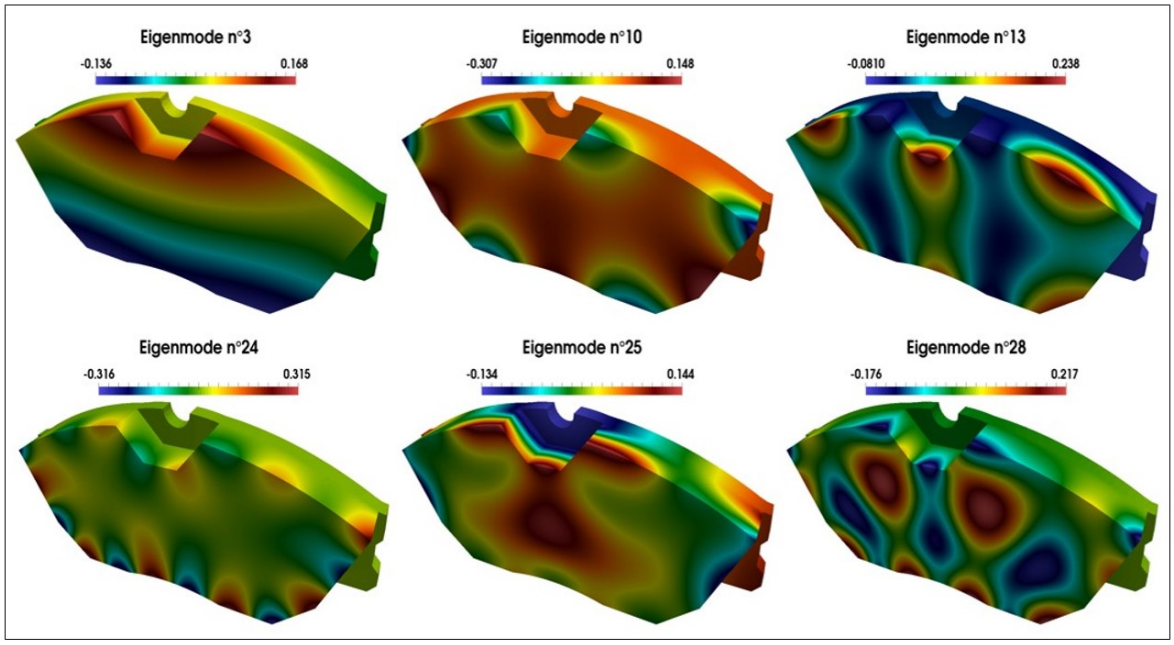


Figure 11: Temperature basis

The reduced modal expression of the thermal problem defined by the equation (10) is then:

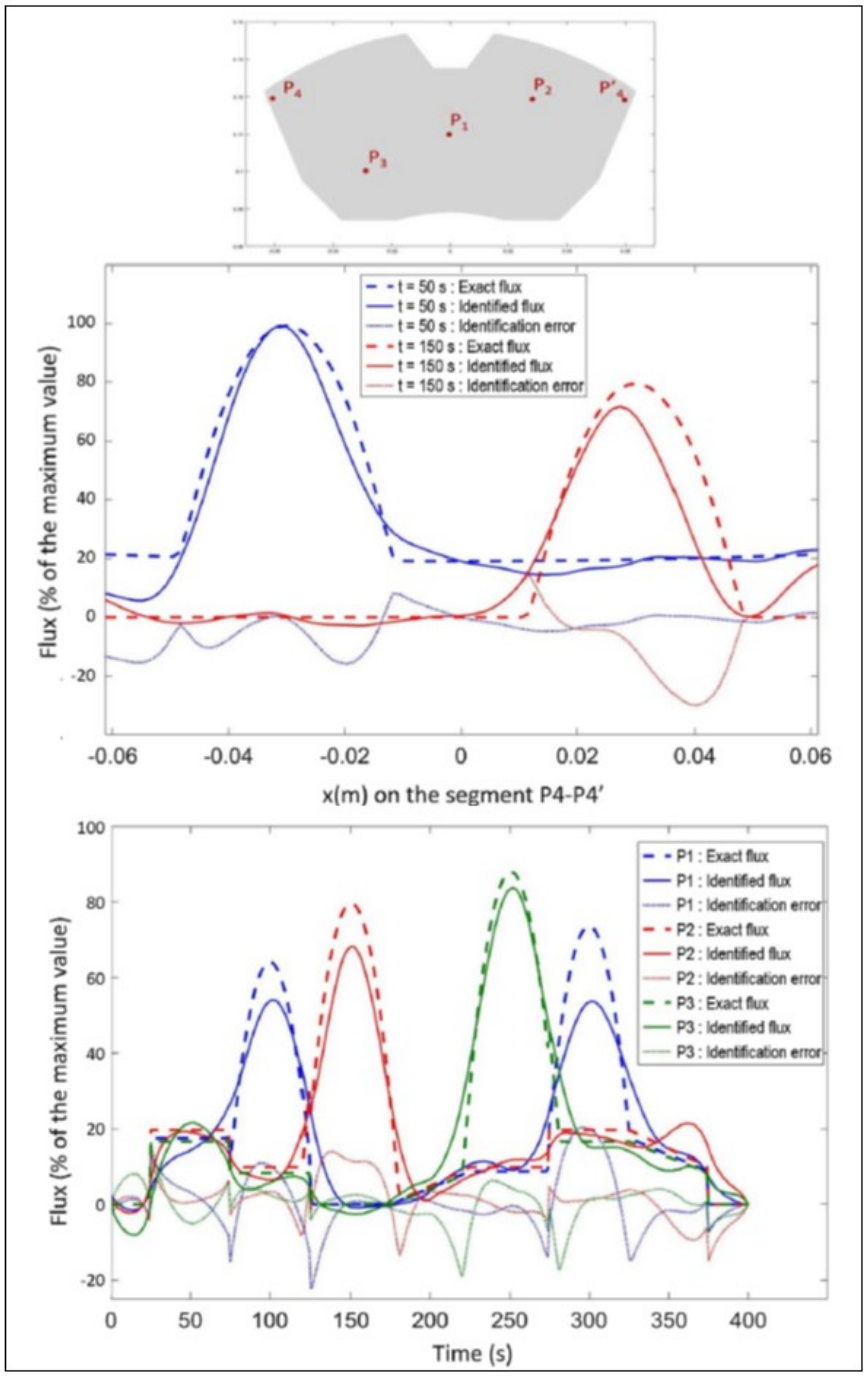
$$\mathbf{L}\dot{\tilde{\mathbf{X}}}^{(\mathbf{T})} = \mathbf{M}\,\tilde{\mathbf{X}}^{(\mathbf{T})} + \mathbf{D}\,\tilde{\mathbf{X}}^{(\boldsymbol{\varphi})} \tag{65}$$

with $\mathbf{D} = \tilde{\mathbf{V}}^{(\mathbf{T})t}\ \mathbf{W}\ \tilde{\mathbf{V}}^{(\boldsymbol{\varphi})}$

6.2.3 space time identification

We thus have a temperature model characterized by a few tens of excitation states of temperature x^T (instead of 67353 degrees of freedom of the initial mesh), to identify a few tens of excitation states of flux x^{φ}, instead of the 5945 degrees of freedom of the surface $\Gamma 1$. The

developed technique uses an iterative method of conjugate gradient descent, for which the gradient is estimated by the adjoint method.
The obtained results (Figures 12 and 13) are satisfactory. It can be noted that no specific regularization technique is used in this study (Tikhonov for example). Indeed, in addition to the natural regularization obtained by using a whole time-domain approach and an iterative method, an additional regularization appears, which is induced by the use of the two reduced bases (one for the thermal problem and another for the heat flux parametrization).



;

Figure 12: Identification results along a segment or versus time

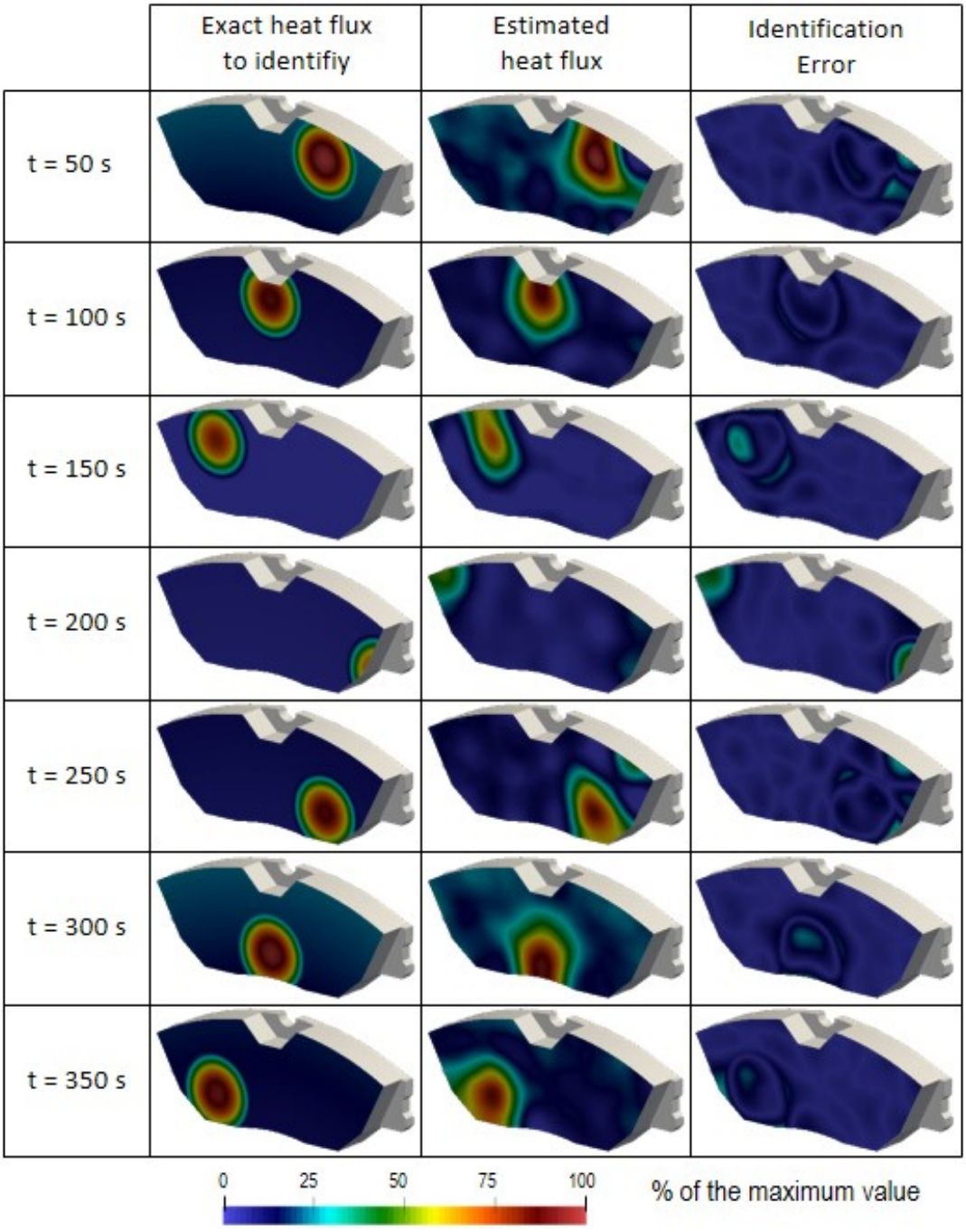


Figure 13: Space-time identification

6.3 On-line indirect thermal measurement in a radiant furnace [3]

6.3.1 The physical problem

Let a heated object on a furnace (Figure 14) in which two radiant tubes dissipate an infra-red radiative heat flux. The power radiated by each tube is driven by the temperature $T_{gas}(t)$ of their intern gas whose value depends on time. The heat exchange between the gas and the tube walls Ω_{tube} is modeled by a global heat exchange coefficient $h_{gas} = 10,000\, W.m^{-2}.K^{-1}$. Given the high temperature level, heat exchange by radiation is preponderant. It is modelled by the radiosity method, which relates the mean flux $\overline{\varphi}_i$ exchanged by patch Ω_i^e to the set of mean temperatures $\overline{T}_j$, with $j \in [1, N_p]$:

$$\forall j \in [1, N_p] \quad \sum_{i=1}^{N_p} \left[\frac{\delta_{ji}}{\varepsilon_i} - \left(\frac{1}{\varepsilon_i} - 1 \right) F_{ji} \right] \overline{\varphi}_i = - \sum_{i=1}^{N_p} \left(\delta_{ji} - F_{ji} \right) \sigma \overline{T}_i^4 \,, \tag{66}$$

where δ_{ji} is the Kronecker delta and F_{ji} are the view factors. This relation (66) can be written in matrix form :

$$\mathbf{A}\, \overline{\varphi} = \mathbf{B}\, \overline{\mathbf{T}}^{\mathbf{4}} \,. \tag{67}$$

The mean flux exchanged by a patch $\overline{\varphi}_j$ expresses as:

$$\overline{\varphi}_j = \sum_{i=1}^{N_p} r_{ji} \overline{T}_i^4 \,, \tag{68}$$

where r_{ji} are the elements of $\mathbf{R}_{rad}\,[N_p,\, N_p] = \mathbf{A}^{-1}\,\mathbf{B}$.

This radian flux is included in the heat equation defined on the solid domains of the scene (wall, tubes, stand, etc) , whicb can be written after s(Figure 14) :

$$\mathbf{C} \frac{d\mathbf{T}}{dt} = [\mathbf{K} + \mathbf{H}]\, \mathbf{T} + \mathbf{U_{cpl}}\, T_{int}\,(\mathbf{T}) + \mathbf{U_0} + \overline{\mathbf{R}}_{rad} \overline{\mathbf{T}}^4 + T_{gas}(t)\ \mathbf{U_{tube}} \,. \tag{69}$$

In this equation:

- Vector $\mathbf{T}$ contains the temperature value at the N discretization points.
- $\mathbf{C}$, $\mathbf{K}$ and $\mathbf{H}$ are $[N \times N]$ symmetric sparse matrices: $\mathbf{C}$ is the thermal inertia matrix, $\mathbf{K}$ the conductivity matrix and $\mathbf{H}$ gathers the different convection terms on Ω_{ext}, Ω_{int} and Ω_{tube}.
- Vector $\mathbf{U_0}$ corresponds to the external known solicitations and $\mathbf{U_{cpl}}$ represents the convective exchange with the air inside the furnace, at temperature that depends on the temperature of all internal surfaces $T_{int}\,(\mathbf{T})$:

$$T_{int}\,(T) = \frac{\int\limits_{\Omega_{int}} h_{int}\ T\ d\Omega}{\int\limits_{\Omega_{int}} h_{int}\ d\Omega} \,. \tag{70}$$

  We obtein after discretization :

$$T_{int}\,(\mathbf{T}) = \mathbf{D}\,\mathbf{T} \,. \tag{71}$$

- Vector $\overline{\mathbf{T}}^4$ of dimension $[N_p]$ contains mean temperatures of every patch Ω_i^e. Radiation matrix $\overline{\mathbf{R}}_{rad}$ $[N \times N_p]$ allots the mean heat flux density from the N_p patches to the N nodes.

$$\overline{\mathbf{T}} = \mathbf{U_R}\,\mathbf{T}\,, \tag{72}$$

- Finally, vector $\mathbf{U_{tube}}$ of dimension $[N]$ stands for the heat source generated by the gas combustion inside the radiant tubes.

6.3.2 Identification and reconstruction of the thermal field

The goal is to recover the whole thermal field of the heated object from a few measurement points (A, B and C on figure 14). The radiant thermal source is first identified via a low order reduced model based on AROMM method (Figure 15).

From this identified temperature, the thermal field is then recovered by direct simulation using a reduced model of higher order which leads to a better precision.

The whole identification procedure lasts less than $5\,000\,s$, which is ten times smaller the duration of the thermal process ($50\,000\,s$). The whole thermal field of the heated object is refreshed every $200\,s$ with an average precision of $\bar{\sigma} = 2.9\,K$, which is below the measurement noise.

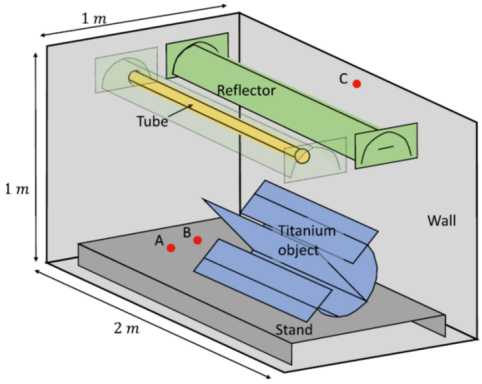


Figure 14: The considered geometry

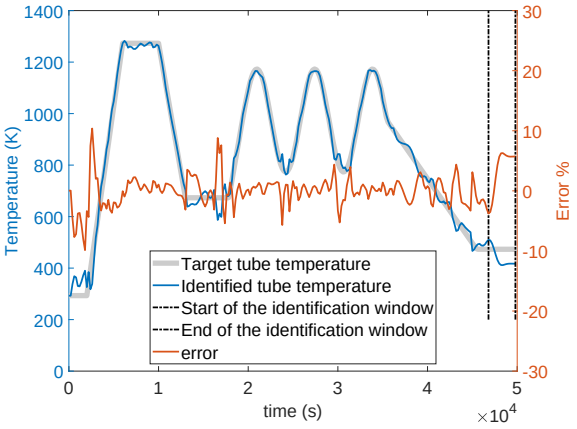


Figure 15: Identification results with $\widetilde{N}_{(rec)} = 20$ modes and $\sigma_N = 5\,K$.

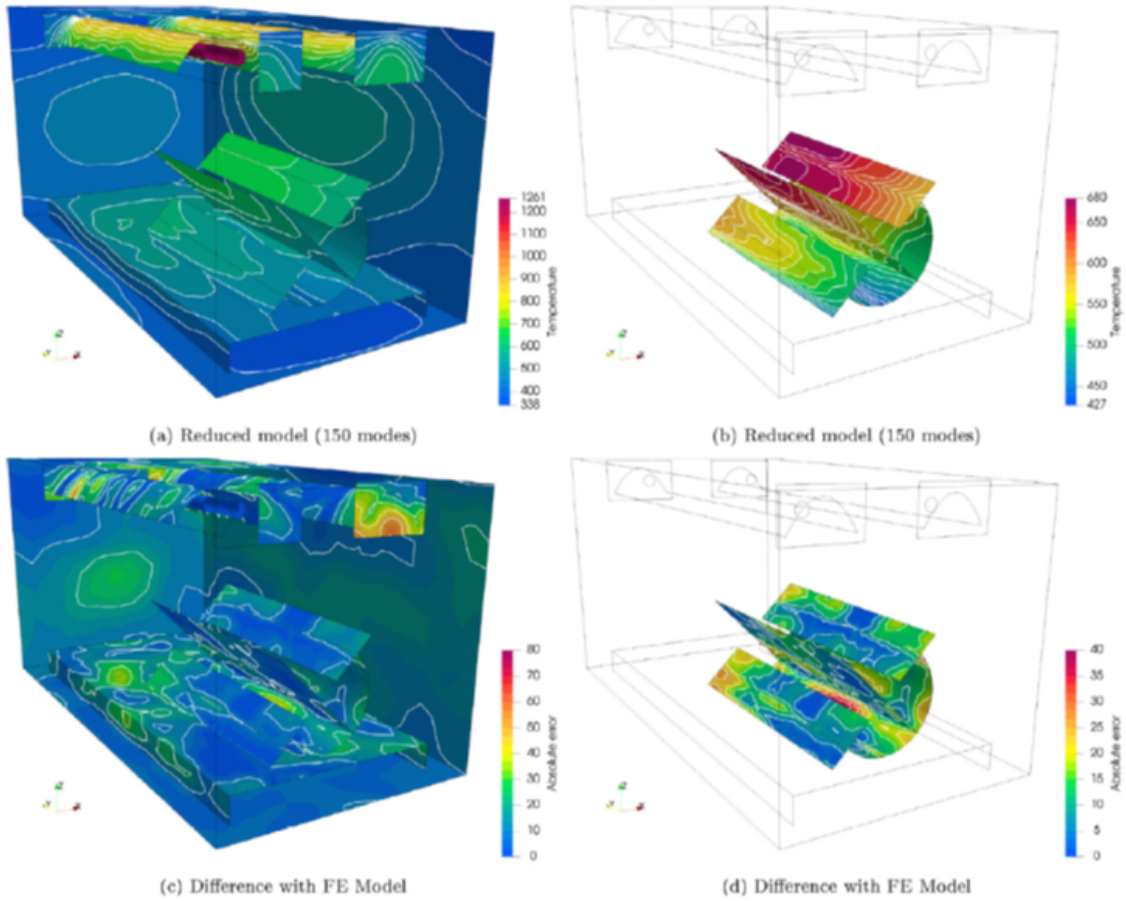


Figure 16: Temperature and error fields at $t = 6\,000\,s$ for reconstruction with $\widetilde{N}_{(rec)} = 150$ modes in the case of identification with $\widetilde{N}_{(id)} = 20$ modes and $\sigma_N = 5\,K$.

References

[1] S. Carmona, Y. Rouizi, O. QuÃ©mÃ©ner, F. Joly, A. Neveu, Estimation of heat flux by using reduced model and the adjoint method. application to a brake disc rotating (2018).

[2] S. Carmona, Y. Rouizi, O. QuÃ©mÃ©ner, Spatio-temporal identification of heat flux density using reduced models. application to a brake pad, International journal of heat and mass transfer 128 (2019) 1048–1063.

[3] B. Gaume, F. Joly, O. Quéméner, Modal reduction for a problem of heat transfer with radiation in an enclosure, International Journal of Heat and Mass Transfer-doi:https://doi.org/10.1016/j.ijheatmasstransfer.2019.07.039.

[4] S. Grosjean, F. Joly, K. Vera, A. nevau, E. Monier-Vinard, Reduction of an electronic card thermal problem by the modal substructuring method, 16th International Hreat Transfer Conference, August 10-15, Beijing, China.

[5] A. Fic, R. Bialecki, A. Kassab, Solving transient nonlinear heat conduction problems by proper orthogonal decomposition and the finite-element method, Numerical Heat Transfer, Part B: Fundamentals 48 (2) (2005) 103–124. arXiv:http://dx.doi.org/10.1080/10407790590935920, doi:10.1080/10407790590935920.

[6] X. Zhang, H. Xiang, A fast meshless method based on proper orthogonal decomposition for the transient heat conduction problems, International Journal of Heat and Mass Transfer 84 (2015) 729 – 739. doi:http://dx.doi.org/10.1016/j.ijheatmasstransfer.2015.01.008.

[7] R. Ghosh, Y. Joshi, Error estimation in pod-based dynamic reduced-order thermal modeling of data centers, International Journal of Heat and Mass Transfer 57 (2) (2013) 698 – 707. doi:http://dx.doi.org/10.1016/j.ijheatmasstransfer.2012.10.013.

[8] A. Sempey, C. Inard, C. Ghiaus, C. Allery, Fast simulation of temperature distribution in air conditioned rooms by using proper orthogonal decomposition, Building and Environment 44 (2) (2009) 280 – 289. doi:http://dx.doi.org/10.1016/j.buildenv.2008.03.004.

[9] J. García, J. Cabeza, A. Rodríguez, Two-dimensional non-linear inverse heat conduction problem based on the singular value decomposition, International Journal of Thermal Sciences 48 (6) (2009) 1081 – 1093. doi:http://dx.doi.org/10.1016/j.ijthermalsci.2008.09.002.

[10] A. Rajabpour, F. Kowsary, V. Esfahanian, Reduction of the computational time and noise filtration in the IHCP by using the proper orthogonal decomposition POD method, International Communications in Heat and Mass Transfer 35 (8) (2008) 1024 – 1031. doi:http://dx.doi.org/10.1016/j.icheatmasstransfer.2008.05.004.

[11] H. Park, M. Sung, Sequential solution of a three-dimensional inverse radiation problem, Computer Methods in Applied Mechanics and Engineering 192 (33) (2003) 3689 – 3704. doi:http://dx.doi.org/10.1016/S0045-7825(03)00370-0.

[12] W. Adamczyk, Z. Ostrowski, Retrieving thermal conductivity of the solid sample using reduced order model inverse approach, International Journal of Numerical Methods for Heat & Fluid Flow 27 (3) (2017) 729–739. arXiv:https://doi.org/10.1108/HFF-05-2016-0206, doi:10.1108/HFF-05-2016-0206.

[13] M. Girault, D. Petit., Identification methods in nonlinear heat conduction. Part I: Model Reduction, International Journal of Heat and Mass Transfer 48 (2005) 105–118.

[14] M. Girault, D. Petit, Identification methods in nonlinear heat conduction. Part II: inverse problem using a reduced model, International Journal of Heat and Mass Transfer 48 (1) (2005) 119 – 133. doi:DOI: 10.1016/j.ijheatmasstransfer.2004.06.033.

[15] E. Videcoq, M. Girault, V. Ayel, C. Romestant, Y. Bertin, On-line thermal regulation of a capillary pumped loop via state feedback control using a low order model, Applied Thermal Engineering 108 (2016) 614 – 627. doi:http://dx.doi.org/10.1016/j.applthermaleng.2016.07.071.

[16] M. Girault, E. Videcoq, D. Petit, Estimation of time-varying heat sources through inversion of a low order model built with the modal identification method from in-situ temperature measurements, International Journal of Heat and Mass Transfer 53 (1) (2010) 206 – 219. doi:http://dx.doi.org/10.1016/j.ijheatmasstransfer.2009.09.040.

[17] K. Bouderbala, H. Nouira, E. Videcoq, M. Girault, D. Petit, MIM, FEM and experimental investigations of the thermal drift in an ultra-high precision set-up for dimensional metrology at the nanometre accuracy level, Applied Thermal Engineering 94 (2016) 491 – 504. doi:http://dx.doi.org/10.1016/j.applthermaleng.2015.09.092.

[18] E. Videcoq, M. Girault, A. Piteau, Thermal control via state feedback using a low order model built from experimental data by the modal identification method,

International Journal of Heat and Mass Transfer 55 (5) (2012) 1679 – 1694. doi:http://dx.doi.org/10.1016/j.ijheatmasstransfer.2011.11.023.

[19] J. Berger, S. Guernouti, M. Woloszyn, F. Chinesta, Proper generalised decomposition for heat and moisture multizone modelling, Energy and Buildings 105 (2015) 334 – 351. doi:http://dx.doi.org/10.1016/j.enbuild.2015.07.021.

[20] J. Berger, W. Mazuroski, N. Mendes, S. Guernouti, M. Woloszyn, 2d whole-building hygrothermal simulation analysis based on a pgd reduced order model, Energy and Buildings 112 (2016) 49 – 61. doi:http://dx.doi.org/10.1016/j.enbuild.2015.11.023.

[21] J. Berger, N. Mendes, An innovative method for the design of high energy performance building envelopes, Applied Energy 190 (2017) 266 – 277. doi:http://dx.doi.org/10.1016/j.apenergy.2016.12.119.

[22] D. González, F. Masson, F. Poulhaon, A. Leygue, E. Cueto, F. Chinesta, Proper generalized decomposition based dynamic data driven inverse identification, Mathematics and Computers in Simulation 82 (9) (2012) 1677 – 1695. doi:http://dx.doi.org/10.1016/j.matcom.2012.04.001.

[23] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, Journal of Research of the National Bureau of Standards 45 (1950) 255–282. doi:10.6028/jres.045.026.

[24] R. Radke, A matlab implementation of the implicitly restarted arnoldi method for solving large-scale eigenvalues problems, Ph.D. thesis, A thesis submitted in partial fullfillment of the requirements for the degree Master of Arts, Rice University, Houston, Texas (1996).

[25] R. Lehoucq, D. Sorensen, C. Yang, ARPACK Users' Guide: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, SIAM e-books, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1998.

[26] J. Sicard, P. Bacot, A. Neveu, Analyse modale des échanges thermiques dans le bâtiment, International Journal of Heat and Mass Transfer 28 (1) (1985) 111 – 123. doi:http://dx.doi.org/10.1016/0017-9310(85)90013-4.

[27] P. Bacot, A. Neveu, J. Sicard, Analyse modale des phénomènes thermiques en régime variable dans le bâtiment, Revue Générale de Thermique 267 (1984) 189–201.

[28] G. Lefebvre, J. Bransier, A. Neveu, Simulation du comportement thermique d'un local par des méthodes numériques d'ordre réduit, Revue Générale de Thermique 302 (1987) 106–114.

[29] J. Salgon, A. Neveu, Application of modal analysis to modelling of thermal bridges in buildings, Energy and Buildings 10 (2) (1987) 109 – 120. doi:http://dx.doi.org/10.1016/0378-7788(87)90013-2.

[30] A. Neveu, K. El-Khoury, B. Flament, Simulation de la conduction non linéaire en régime variable: décomposition sur les modes de branche, International Journal of Thermal Sciences 38 (4) (1999) 289 – 304. doi:http://dx.doi.org/10.1016/S1290-0729(99)80095-7.

[31] O. Quéméner, A. Neveu, E. Videcoq, A specific reduction method for the branch modal formulation: Application to a highly non-linear configuration, International Journal of Thermal Sciences 46 (9) (2007) 890 – 907. doi:DOI: 10.1016/j.ijthermalsci.2006.11.011.

[32] E. Videcoq, O. Quéméner, W. Nehme, A. Neveu, Real time heat sources identification by a branch eigenmodes reduced model, in: 6th International Conference on Inverse Problems in Engineering, Theory and Practice, Dourdan (France), and in Journal of Physics: Conference Series 135 (freely available on line, paper n 012101), 2008.

[33] E. Videcoq, O. Quéméner, M. Lazard, A. Neveu, Heat source identification and on-line temperature control by a branch eigenmodes reduced model, International Journal of Heat and Mass Transfer 51 (19-20) (2008) 4743 – 4752. doi:DOI: 10.1016/j.ijheatmasstransfer.2008.02.029.

[34] E. Videcoq, M. Lazard, O. Quéméner, A. Neveu, Online temperature prediction using a branch eigenmode reduced model applied to cutting process, Numerical Heat Transfer, Part A: Applications 55 (7) (2009) 683–705. arXiv:http://dx.doi.org/10.1080/10407780902821490, doi:10.1080/10407780902821490.

[35] F. Joly, O. Quéméner, A. Neveu, Modal reduction of an advection-diffusion model using a branch basis, Numerical Heat Transfer, Part B: Fundamentals 53 (5) (2008) 466–485. doi:10.1080/10407790701849550.

[36] O. Quéméner, F. Joly, A. Neveu, On-line heat flux identification from a rotating disk at variable speed, International Journal of Heat and Mass Transfer 53 (7) (2010) 1529 – 1541. doi:http://dx.doi.org/10.1016/j.ijheatmasstransfer.2009.11.032.

[37] P. Laffay, O. Quéméner, A. Neveu, Developing a method for coupling branch modal models, International Journal of Thermal Sciences 48 (6) (2009) 1060 – 1067. doi:http://dx.doi.org/10.1016/j.ijthermalsci.2008.11.002.

[38] P. O. Laffay, O. Quéméner, A. Neveu, B. Elhajjar, The modal substructuring method: An efficient technique for large-size numerical simulations, Numerical Heat Transfer, Part B: Fundamentals 60 (4) (2011) 278–304. doi:10.1080/10407790.2011.609113.

[39] S. Marshall, An approximation method for reducing the order of linear system, control (1966) 642–643.

[40] O. Quéméner, J. Battaglia, A. Neveu, Résolution d'un problème inverse par utilisation d'un modèle réduit modal. application au frottement d'un pion sur un disque en rotation, International Journal of Thermal Sciences 42 (4) (2003) 361 – 378. doi:http://dx.doi.org/10.1016/S1290-0729(02)00037-6.

[41] A. Oulefki, Réduction des modèles thermiques par amalgame modal, Ph.D. thesis, Ecole Nationale de Ponts et Chaussées (1993).

[42] G. Benjamin, Réduction d'un problème d'auto-rayonnement par modes de branche : application aux échanges thermiques dans un domaine multi-enceintes, Ph.D. thesis, Paris Saclay (2016).

[43] N. Brou, Modélisation et commande d'un système de cogénération utilisant des énergies renouvelables pour le bâtiment, Ph.D. thesis, Paris Saclay (2015).

[44] O. Quéméner, F. Joly, A. Neveu, The generalized amalgam method for modal reduction, International Journal of Heat and Mass Transfer 55 (4) (2012) 1197 – 1207. doi:http://dx.doi.org/10.1016/j.ijheatmasstransfer.2011.09.043.

Lecture 8 : Optimization tools dedicated to function estimation in inverse heat transfer problems

Y. Favennec
LTeN – Université de Nantes – France
yann.favennec@univ-nantes.fr
www.univ-nantes.fr/yann-favennec
+33 (0)240 683 138

Abstract. This lecture presents some commonly-used numerical algorithms devoted to optimization, that is maximizing or, more often minimizing a given function of several variables. The goal is function estimation. At first, some general mathematical tools are presented. Some gradient-free optimization algorithms are presented and then some gradient-type methods are pointed out with pros and cons for each method. The gradient of the function to be minimized is presented according to three distinct methods: finite difference, forward differentiation and the use of the additional adjoint-state problem. The last part presents some practical studies where some tricks are given, along with some numerical results.

Keywords. optimization, convexity, zero-order method, deterministic method, stochastic method, gradient-type method, conjugate gradient, BFGS, Gauss–Newton, Levenberg–Marquardt, gradients, direct differentiation, adjoint-state

⋆ ⋆ ⋆

Contents

⋆ ⋆ ⋆

1 Introduction

This lecture is devoted to the solution of inverse problems in heat transfer, specifically when function are to be recovered. Usually, such problems are non-linear and may fall into the category of large-scale inverse problems, so that specific optimization tools are to be developed.
The lecture first presents some basic examples of IHCP (Inverse Heat Conduction Problems) and points out the distinction between estimation of parameters on the one hand, and functions in the other hand. Indeed, as a simple example, we have the distinction between estimating $i)$ λ as a parameter, $ii)$ $\lambda(\boldsymbol{x})$ as a function of the space $\boldsymbol{x}$ ($\boldsymbol{x} = (x_1, x_2)^t$ for instance) and, $iii)$ $\lambda(T)$ as a function of the state T.

The lecture then presents the most usual optimization tools for the solution of different kinds of inverse problems. It first gives notions on the functional to be minimized, and convexity. It gives definitions of constraints (equality and inequality) added to the functional to be minimized, the added constraints being related to either the state or the parameter/functional.
Then, before tackling the detailed iterative optimization algorithms, the most usual stopping criteria are presented.
Zero-order, first-order and quasi-second order optimization methods are briefly presented with pros and cons for each of them.
Concerning zero order methods, both deterministic and stochatic methods are very briefly presented with some specific examples (Simplex, PSO, and GA).
Within the frame of first-order methods, one presents the steepest-descent method with and without line-search, then the conjugate gradient method for quadratic and arbitrary functions.
Some quasi-Newton algorithms are then presented: the BFGS, the Gauss–Newton and the Levenberg–Marquardt methods.
A comparison is given in terms of gradient needed for all previously presented methods along with the convergence rate, if possible.
The next part presents the computation of the functional gradient: through the finite difference method, through the direct differentiation of the PDEs (partial differential equations), and through the use of the adjoint-state problem. Several ways to access the adjoint-state problems are given. A comparison of gradient computations is given through examples to emphasize the differences.
Note that this lecture has been prepared with some well-known books such as [1, 2, 3, 4]. These books being considered as "standard" popular books, some parts of this lecture are taken from these references.
Note also that this lecture is being continuously improved, starting from its very first version in 2005 [5].

2 Estimation in heat transfer – Optimization

2.1 Parameter and function estimation

The modeling of a physical system is based on several requirements. In addition to the physical modeling equations that include some physical parameters (e.g. conductivity coefficients), the initial state and the sources are also to be known if the physical problem is to be solved. If all these data is known, then the so-called *direct problem* – or *forward problem* – can be solved.

Now, if some of the previously expressed quantities are missing, the physical problem cannot be solved any longer, but some inversion procedure may evaluate the missing quantity, fitting the model output with some real ones (i.e. obtained through experiments). The evaluation of such missing quantities needs an *inverse problem* – or a *backward problem* – to be solved.

Depending on the nature of the missing quantity, the estimation is performed on parameters or on functions.

These last years, a debate took place within the heat transfer community about the difference and the meaning of, on the one hand, *parameter identification* and, on the other hand, *function estimation*. According to the author, both are very different, though some similarities exist between both of them.

Let us work on following examples of physical properties estimation to back up our methodology.

i) If a single material conductivity λ is to be identified, then the problem clearly belongs to the category of *parameter estimation*. In such a case, the number of unknowns (the parameters) is very low: only one for a uniform isotropic medium, and only six at maximum for a uniform orthotropic medium. Due to the low dimensionality of the inverse problem, any optimizer can be used (either gradient-free or gradient-type). Moreover, such problems are likely to be well-posed, and the use of regularization tools may not be necessary. These parameter estimation problems are not difficult both from mathematical and computational points of view. The same comments can be drawn if different non-varying thermal

conductivities are to be estimated in different locations (for example dealing with the case of a multi-layer medium).

ii) If several physical properties are to be estimated, for example a thermal conductivity λ [W m^{-1} K^{-1}]), a heat capacity C_p [J K^{-1}]), and a convective heat transfer coefficient h [W m^{-2} K^{-1}]), then we consider a *collection* of elements that can be put together into a vector, such that classical optimizers can solve this *parameter estimation* problem. However, taking a norm of such a vector would not make any sense in a physical point of view. This is one of the reasons why some priors are used (in this specific case λ^0, C_p^0 and h^0), and the estimation is performed on adimensionalized parameters (in this specific case $\tilde{\lambda} = \lambda/\lambda^0$, $\tilde{C}_p = C_p/C_p^0$ and $\tilde{h} = h/h^0$). Doing so, norms (on the collection of adimensionalized parameters) are understandable by both mathematicians and physicists. Note that another reason why it is preferable to adimensionalize parameters is that it usually slightly attenuates the ill-posed character of the inverse problem, and thus the process of adimensionalization can be seen, somehow, as the very first regularization tool.

iii) If a physical property now depends continuously on the state, (e.g. temperature-dependent conductivity $\lambda(T)$), then one may think that the problem of conductivity estimation falls into the category of function estimation. However, a parameterization of this function is anyway necessary to use numerical algorithms, and the type of parameterization can make the difference between parameter and function estimation. If – for example – a polynomial expansion is used, say $\lambda(T) \approx \sum_i^N \alpha_i T^i$, then the collection of the N coefficients α_i is to be estimated, and, therefore, such a problem eventually falls into the category of a *parameter estimation* problem (the parameters are the polynomial coefficients). Moreover, because the number of unknowns is likely to be low (say less than ten), the choice of the optimizer does not matter much. (Note however that this choice of polynomial expansion is unlikely to be a good candidate for the parameterization; the one presented in the following item iv) is likely to be much better.)

iv) If a space-dependent physical property is to be estimated, for example a thermal conductivity $\lambda(\boldsymbol{x})$, then the estimation is performed on a function. As in the previous case, a parameterization of this function is anyway necessary to use any numerical algorithm. Building a basis $\{\xi_i\}_{i=1}^N$ and using it to project the function, i.e. with $\lambda(\boldsymbol{x}) = \sum \lambda_i \xi_i(T)$, the estimation in the end is performed on discrete parameters $\{\lambda_i\}_{i=1}^N$, all of these having the same unit, say [W m^{-1} K^{-1}]. At this stage, one may think we face again a parameter estimation problem. However, most often, the function has to satisfy some regularity properties. For example, the conductivity is finite and varies continuously in space, so $\lambda(\boldsymbol{x}) \in H^1(\mathcal{D}) = \{\lambda \in L^2(\mathcal{D}), \|\lambda\| \in L^2(\mathcal{D})\}$. Because such a regularity property is be satisfied, this problem falls into the category of a *function estimation* problem, and specific regularization tools are to be used to enforce the function to satisfy these constraints of regularity. Added to that, the dimension of the discrete unknown, N, is very likely to be big. (As an example, a property defined in a cube discretized with only 100 voxels per side gives 10^6 unknowns.) Therefore, some specific optimization algorithms have been designed to cope with such high dimensions.

It could be seen from previous examples how function estimation is different from parameter estimation. Main differences between both of them come from, on the one hand, the regularity of the functions to be estimated, and, on the other hand, the high dimensionality of the optimization problem due to the parameterization. The regularity issue demands specific regularization tools and a special care on the parameterization, and the high dimensionality demands powerful optimization algorithms.

2.2 The function to be minimized

In an inversion process, one usually minimizes some errors between some experimental data (say u_d) and related model data (say u). The cost function (also called somewhere discrepancy function or objective

function) is often expressed as the square of a norm of the difference between u and u_d. The most often, one uses the $L_2(\cdot)$ norm if some "quasi-"continuous u and especially u_d are available (i.e. $\|u-u_d\|^2_{L_2(\mathcal{S})} = \int_{\mathcal{S}}(u-u_d)^2\,\mathrm{d}\boldsymbol{x}$ but, when data u_d is given only on specific locations (in space and/or time) , then the squared euclidean norm is to be used: $\|u-u_d\|^2_2 := \sum_i (u(\boldsymbol{x}_i)-u_d(\boldsymbol{x}_i))^2 = \int_{\mathcal{S}}\delta^j_i(u-u_d)^2\,\mathrm{d}\boldsymbol{x}$ where $\delta^j_i = \delta(\boldsymbol{x}^i-\boldsymbol{x}^j)$. Often, some function of the state and of the measure are used, for instance state derivation, integration, weighted summation, etc. Moreover, some selection process is, most of the time considered. So, in order to write down a general form for the cost function to be minimized, we use :

$$\mathcal{J}(u) = \|u-u_d\|^2_{\mathcal{X}} \tag{1}$$

without specifying any choice for the norm on $\mathcal{X}$ at this early stage. Though the cost function is explicitly given in terms of the state u, the cost function is actually to be minimized with respect to what it is searched, *i.e.* the parameters ψ. Hence we write the equality (by definition):

$$j(\psi) := \mathcal{J}(u,\psi) \tag{2}$$

where the function j is often called the reduced cost function, as opposed to $\mathcal{J}$ which is the calculated cost function. One actually computes the cost function in terms of the state (by eq. (1) for instance), but the cost function is to be minimized with respect to another quantity, say ψ.

2.3 Elements of minimization

The function denoted j in eq. (2) is defined on $\mathcal{K}$ with values in $\mathbb{R}$. $\mathcal{K}$ is a set of admissible elements of the problem. In some cases, $\mathcal{K}$ defines some constraints on the parameters or functions. The minimization problem is written as:

$$\inf_{\phi\in\mathcal{K}\subset\mathcal{V}} j(\phi).$$

According to [1], if the notation "inf" is used for a minimization problem, it means that one does not know, *a priori*, if the minimum is obtained, *i.e.* if there exists $\phi\in\mathcal{K}$ such that

$$j(\phi) = \inf_{\psi\in\mathcal{K}\subset\mathcal{V}} j(\psi).$$

For indicating that the minimum is obtained, one should prefer the notations

$$\phi = \arg\min_{\psi\in\mathcal{K}\subset\mathcal{V}} j(\psi) \text{ and } j(\phi) = \min_{\psi\in\mathcal{K}\subset\mathcal{V}} j(\psi)$$

Let us now recall basic definitions needed for mathematical optimization [1]:

Definition 1. *ψ is a local minimum of j on $\mathcal{K}$ if and only if*

$$\psi\in\mathcal{K} \text{ and } \exists\delta>0,\ \forall\phi\in\mathcal{K}, \|\phi-\psi\|<\delta \ \rightarrow j(\phi)\geq j(\psi).$$

Definition 2. *ψ is a global minimum of j on $\mathcal{K}$ if and only if*

$$\psi\in\mathcal{K} \text{ and } j(\phi)\geq j(\psi)\ \forall\phi\in\mathcal{K}.$$

Definition 3. *A minimizing series of j in $\mathcal{K}$ is a series $(\psi^n)_{n\in\mathbb{N}}$ such that*

$$\psi^n\in\mathcal{K}\ \forall n \text{ and } \lim_{n\rightarrow+\infty} j(\psi^n) = \min_{\phi\in\mathcal{K}} j(\phi).$$

Definition 4. *a set $\mathcal{K}\in\mathcal{V}$ is convex if, for all $\psi,\phi\in\mathcal{K}$ and $\forall\theta\in[0,1]$, the element $(\theta\psi+(1-\theta)\phi)$ is in $\mathcal{K}$ (see figure 1).*

Definition 5. *A function j is said to be convex when defined on a non-null convex set $\mathcal{K}\in\mathcal{V}$ with values in $\mathbb{R}$ if and only if*

$$j(\theta\psi+(1-\theta)\phi)\leq\theta j(\psi)+(1-\theta)j(\phi)\ \forall\psi,\phi\in\mathcal{K},\ \forall\theta\in[0,1].$$

Moreover, j is said to be strictly convex if the inequality is strict when $\psi\neq\phi$ and $\theta\in\,]0,1[$ (see fig. 2).

Ending, if j if a convex function on $\mathcal{K}$, the local minimum of j on $\mathcal{K}$ is the global minimum on $\mathcal{K}$.

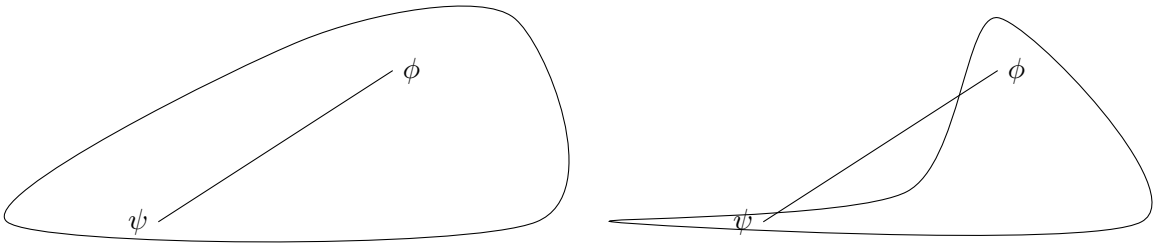


Figure 1: Convex and non-convex domaine $\mathcal{K}$

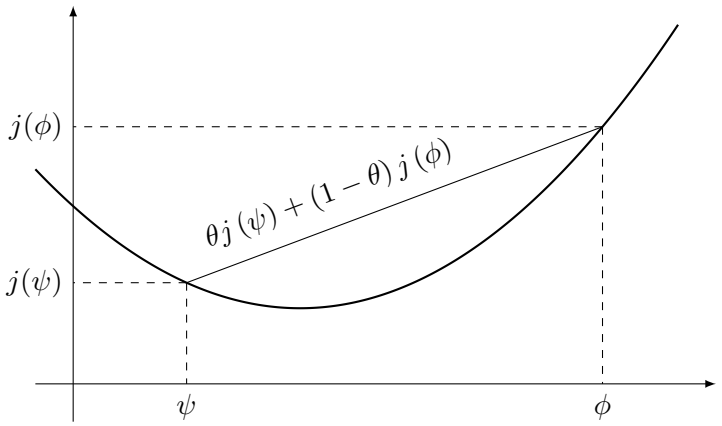


Figure 2: Convex function $j(\cdot)$

2.4 Optimality conditions

For convex functions, there is no difference between local minima and global minimum. In the following we are more interested in minimizing a function without specifying whether the minimum is local or global. It will be seen in next sections that some gradient-free optimization algorithms may find the global minimum even if the cost function contains local minima.

Let us derive here the minimization necessary and sufficient conditions. These conditions use the first-order derivatives (order-1 condition), and second-order derivatives (order-2 condition) on the cost function j. Using gradient-type algorithms, the first–order condition is to be reached, while the second-order condition requires assuming a local convexity hypothesis, and then make a distinction between minima, maxima and optima.

Let us assume that $j(\psi)$ is continuous and has continuous first partial derivatives $(\partial j/\partial\psi_i)(\psi)$ and second partial derivatives $(\partial^2 j/\partial\psi_i\partial\psi_j)(\psi)$. Then a *necessary condition* for $\bar{\psi}$ to be a minimum of j (at least locally) is that:

i) $\bar{\psi}$ is a stationary point, i.e. $\nabla j(\bar{\psi}) = 0$,

ii) the Hessian $\nabla^2 j(\bar{\psi}) = \left(\partial^2 j/\partial\psi_i\partial\psi_j\right)(\bar{\psi})$ is a positive semi-definite matrix, i.e. $\forall y \in \mathbb{R}^n$, $\left(\nabla^2 j(\bar{\psi})y, y\right) \geq 0$ where $(.,.)$ is a scalar product in $\mathbb{R}^n$ (we have $\dim(\psi) = n$).

A point $\bar{\psi}$ which satisfies condition item i) is called a *stationary point*. It is important to point out that stationarity is not a sufficient condition for local optimality. For instance the point of inflexion for cubic functions would satisfy the condition i), while there is no optimum. Hence the Hessian is not positive definite but merely positive semi-definite.

The *sufficient condition* for $\bar{\psi}$ to be a minimum of j (at least locally) is that

i) $\bar{\psi}$ is a stationary point, i.e. $\nabla j(\bar{\psi}) = 0$,

ii) the Hessian $\nabla^2 j(\bar{\psi}) = \left(\partial^2 j/\partial\psi_i \partial\psi_j\right)(\bar{\psi})$ is a positive definite matrix, i.e. $\forall y \in \mathbb{R}^n$, $y \neq 0$, $\left(\nabla^2 j(\bar{u})y, y\right) > 0$.

We remark that the condition item ii) amounts to assuming that j is strictly convex in the neighbourhood of $\bar{\psi}$.

2.5 Stopping criteria

Since the convergence of the iterative algorithms is, in general, not finite, a stopping criterion must be applied. Here below are given some commonly used criteria. We denote ψ^p the vector parameter ψ at the optimization iteration p.

$$\|\nabla j(\psi^p)\| \leq \varepsilon; \tag{3}$$
$$\left|j(\psi^p) - j(\psi^{p-1})\right| \leq \varepsilon; \tag{4}$$
$$\left\|\psi^p - \psi^{p-1}\right\| \leq \varepsilon; \tag{5}$$
$$j(\psi^p) \leq \boldsymbol{v} \tag{6}$$

For each of the above criteria, it may be judicious to demand that the test is satisfied over several successive iterations. The three first above-presented criteria are convergence criteria applied on the cost function gradient, on the cost function evolution, or on the parameter evolutions. These criteria are very commonly-used when dealing with optimization and optimal control problems.

The last criterion is, in one sense, more specific to inverse problems: when the cost function reaches a critical value that depends on the variance of measurement errors, then the optimization algorithm should stop [6, 7, 8]. It can be shown that the consequence of lowering the cost function below $\boldsymbol{v}$ affects the result in dramatically highliting its inherent noise. This criterion is the "maximum discrepancy principle".

Often, the maximum discrepancy principle eq. (6) is used together with the other criteria and also with a maximum number of iterations.

2.6 Classification of optimization methods

The solution of the optimization problem may be performed in numbers of ways. Among numerous methods found in the litterature, the classification of methods given below (see fig. 3) is based on our experience. First, one can distinguish gradient-free methods from methods relying on gradients. Among gradient-free methods, there are those deterministic and those stochastic (the latter introducing random in the search of the optimum). Among gradient-based methods, one can distinguish between first and second-order methods, and those in between.

3 Zero-order n–dimensional optimization algorithms

Zero-order methods, also called "derivative-free optimization" (DFO) or "gradient-free methods" are based on a global vision of the cost function value j on the search space. The main interest of using such methods is when the cost function gradient is not available, or when the cost gradient is not easy to compute, or when the cost function presents local minima. There is an increasing number of computation tools to solve optimization problems with no gradient [9]. In the sequel, we restrict our-self in very briefly presenting, among the enormous number of existing methods, one deterministic algoritm which is the so-called simplex method, and one probabilistic method which is the particle swarm optimization method.

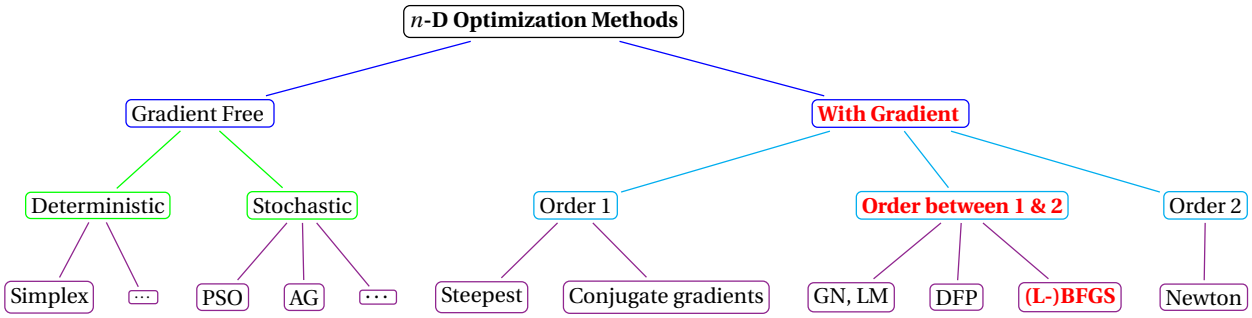


Figure 3: Classification of optimization methods. PSO stands for "partcule swarm optimization", AG stands for "genetic algorithm", GN stands for "Gauss–Newton", and LM stands for Levenberg–Marquardt.

3.1 Simplex

We present here the Nelder-Mead simplex method (1965). This method is popular and simple to code. Moreover, there exists a large number of freeware that can be used to minimize a function using such an algorithm. Let a simplex $\mathscr{S}_0$ be a set of $n+1$ "points" linearly independent ($n = \dim \psi$) with $\mathscr{S}_0 = \{\psi^I, I = 1, \ldots, n+1\}$. One iteration of the simplex optimization algorithm consists in generating a new simplex closer to the minimum eliminating the point with the higher cost function value. The basic operations of $n = 2$ are given in fig. 4: let $\bar{\psi}$ the isobarycenter of $\{\psi^I, I = 1, \ldots, n, \}$ (without $\psi^h = \arg_{I=1,\ldots,n} \max j(\psi^I)$), let the ordering so that

$$j(\psi^1) \leq j(\psi^2) \leq \ldots \leq j(\psi^{n+1})$$

and let $\psi^\ell = \arg_{I=1,\ldots,n} \min j(\psi^I)$. At each iteration, the simplex improvement is performed in three steps:

1. [Reflection] One builds ψ^R symmetry of ψ^h with respect to the segment $[\bar{\psi}, \psi^\ell]$. According to the value of the cost $j(u^R)$ with respect to $j(\psi^\ell)$, the parametric space is then extended (step 2), or contracted (step 3);

2. [Extension] if $j(\psi^R) < j(\psi^\ell)$, one searches a new point in the same direction. The point ψ^E is such that $\psi^E = \gamma\psi^R + (1-\gamma)\bar{\psi}$ with $\gamma > 1$. If $j(\psi^E) < j(\psi^R)$, ψ^h is replaced by ψ^R, otherwise ψ^h is replaced by ψ^E;

3. [Contraction] If $j(\psi^R) > j(\psi^\ell)$, the point ψ^C such that $\psi^C = \gamma\psi^h + (1-\gamma)\bar{\psi}$, $\gamma \in]0, 1[$ is created. If $j(\psi^C) < j(\psi^R)$, ψ^h is replaced by ψ^C otherwise the simplex is contracted (inside contraction) in all directions replacing $\forall I \neq L\ \psi^I$ by $(\psi^I + \psi^\ell)/2$.

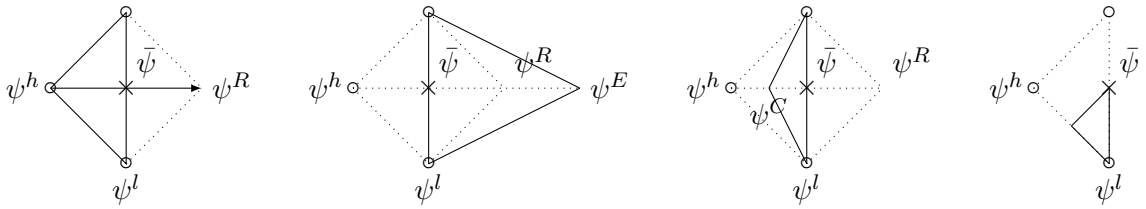


Figure 4: Basic operations on a simplex for $n = 2$. From left to right: reflection, expansion, contraction, and inside contraction.

3.2 PSO

The particle swarm optimization is a stochastic algorithm described by Kennedy and Eberhart in 1995. One considers an initial set of individuals (particles) located randomly. Each particle moves within the space $\mathcal{K}$ interacting with other particles on their best locations. From this information, the particle shall change its position ψ^i and its velocity $\delta\psi^i$. The general formulation for this behavior is given by:

$$\begin{aligned} \delta\psi^i &= \chi\delta\psi^i + \lambda_1 \text{rand}_1 \left(\phi^g - \psi^i\right) + \lambda_2 \text{rand}_2 \left(\phi^i - \psi^i\right) \\ \psi^i &= \psi^i + \delta\psi^i \end{aligned} \tag{7}$$

where ψ^i is the position of the particle i, $\delta\psi^i$ is its velocity, ϕ^g is the best position obtained in its neighborhood, and ϕ^i is its best position (see fig. 5). χ, λ_1 and λ_2 are some coefficients weighting the three directions of the particule [9]:

- how much the particle trusts itself now;
- how much it trusts its experience;
- how much it trusts its neighbours.

Next, rand_1 and rand_2 are random variables following a uniform distribution in $[0, 1]$. There are several configuration parameters for the method, see [10]:

- swarm size, usually between 20 and 30;
- initialization of both the position of the particles and their velocity $\sim \mathcal{U}[0, 1]$;
- neighborhood topology such that a particule communicates with only some other particles;
- inertial factor χ which defines the exploration capacity of the particules;
- confidence coefficients λ_1 and λ_2 which are constriction coefficients;
- stopping criterion which is usually the maximum of iterations, or the critical value of the cost function $j(\psi)$.

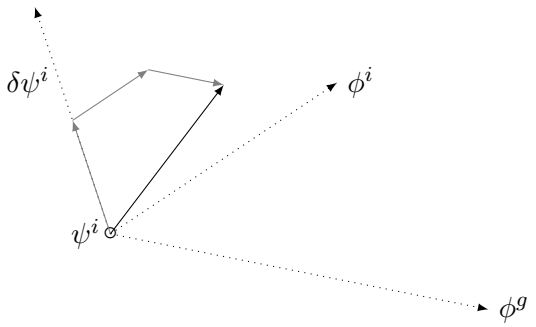


Figure 5: PSO algorithm: a particle displacement.

Usually, a circular neighborhood topology is used, along with $\chi = 0.72$ and $\lambda_1 = \lambda_2 = 1.46$. A large number of free software are available, see for instance [10].

4 One-dimensional unconstrained opimization – line search algorithm

In order to find the optimum of a function j of n variables, we shall describe in section 5 a number of iterative methods which require, at each step, the solution of an optimization problem in one single variable, of the type:

$$\text{Find } \bar{\alpha} = \arg\min_{\alpha>0} g(\alpha) = j\left(\psi^p + \alpha d^p\right), \tag{8}$$

where $\psi^p = (\psi_1^p \dots \psi_n^p)^t$ is the obtained point at iteration p and where $d^p = (d_1^p \dots d_n^p)^t$ is the direction of descent (see section 5). As a matter of fact we have the problem of finding the optimum of the function j, starting from the guess ψ^0 in the direction of descent d^0. Since this problem must be solved a great number of times, it is important to design efficient algorithms that deal with it.In any case, one has to keep in mind that the main objective is not to solve eq. (8) but to find the minimum of $j(\psi)$. Thus one has to design efficient tools for the one-dimensional algorithm that finds the minimum of $g(\alpha)$, or approach it, in a not so expensive way. Note that we always assume that $g'(0) = (\nabla j(\psi^p), d^p) < 0$, which means that d^p is indeed a descent direction.

4.1 The dichotomy method

This method halves, at each step, the length of the interval which contains the minimum, by computing the function g in two new points. By carrying out n computations of the function g, the length of the initial interval $[a^0, b^0]$ is reduced in a proportion of $2^{(n-3)/2}$. The general procedure is the following: starting from the interval $[a^0, b^0]$, and taking the midpoint $c^0 = \left(a^0 + b^0\right)/2$, and the two points $d^0 = \left(a^0 + c^0\right)/2$, and $e^0 = \left(c^0 + b^0\right)/2$, one obtains five equidistant points of length $\delta^0 = (b^0 - a^0)/4$; computing the cost function values at these points, two of the four sub-intervals may be eliminated, while the two adjacent sub-intervals remain; the same procedure is repeated within the selected interval $[a^1, b^1]$, and so on. Since the step length is divided by 2 at each iteration, the dichotomy method converges linearly to the minimum [2].

4.2 The Newton–Raphson method

Let us assume that the function $g(\alpha)$ is twice continuously differentiable. The search for a minimum of $g(\alpha)$ is carried out by looking for a stationary point, *i.e.* $\bar{\alpha}$ satisfying the possibly nonlinear relationship $g'(\bar{\alpha}) = 0$. If α^q is the point obtained at stage q, then the function $g'(\alpha)$ is approximated by its tangent, and the next point α^{q+1} is chosen to be at the intersection of this tangent with the zero-ordinate axis. The relationship to pass from one step to the next comes from $g'\left(\alpha^{q+1}\right) = g'\left(\alpha^q\right) + g''\left(\alpha^q\right) \times \left(\alpha^{q+1} - \alpha^q\right) = 0$ which gives:

$$\alpha^{q+1} = \alpha^q - \frac{g'\left(\alpha^q\right)}{g''\left(\alpha^q\right)}. \tag{9}$$

It is of interest that this method has the property of finite convergence when applied to quadratic functions. This is an interesting feature because any function which is sufficiently regular (at least twice continuously differentiable) behaves as a quadratic function near the optimum [2]. On the other hand, the main drawback of this method is that it requires the computation of the first and of the second derivative of g at each stage. That is the reason why the secant method (next section) is also widely used, especially when there is no way for computing the second order derivative, or when the exact second derivative is complicated to compute or too time consuming.

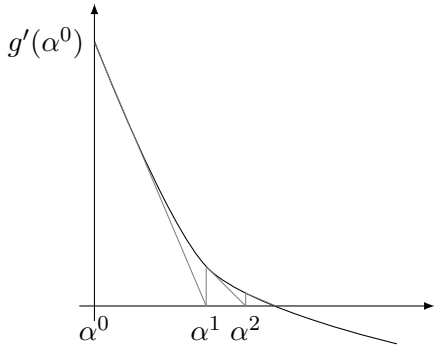


Figure 6: Schematic representation of the Newton–Raphson line search method.

4.3 The secant method

The second-order derivative $g''(\alpha)$ is approximated by finite differences so that the Newton–Raphson's equation initially given by eq. (9) becomes eq. (10):

$$\alpha^{q+1} \approxeq \alpha^q - g'(\alpha^q)\frac{\alpha^q - \alpha^{q-1}}{g'(\alpha^q) - g'(\alpha^{q-1})}. \tag{10}$$

This method is the so-called *secant method*. Applied to the search of $g'(\alpha) = 0$, this method consists in searching the intersection between the zero-ordinate axis and the straight line passing by the points $\left[\alpha^{q-1}, g'(\alpha^{q-1})\right]$ and $[\alpha^q, g'(\alpha^q)]$.

4.4 The quadratic interpolation

By comparison of those of sections 4.2 and 4.3, this method has the advantage of not requiring the computation of first or second order derivatives of the function . Let three points $\alpha_1 \leq \alpha_2 \leq \alpha_3$ such that $g(\alpha_1) \geq g(\alpha_2) \leq g(\alpha_3)$ and let us approximate the function g on the related interval by a quadratic function $\tilde{g}$ with the same values as those of g at the points α_1, α_2 and α_3. The minimum of $\tilde{g}$ is obtained at the new point α_4 satisfying:

$$\alpha_4 = \frac{1}{2}\,\frac{r_{23}g(\alpha_1) + r_{31}g(\alpha_2) + r_{12}g(\alpha_3)}{s_{23}g(\alpha_1) + s_{31}g(\alpha_2) + s_{12}g(\alpha_3)}, \tag{11}$$

where $r_{ij} = \alpha_i^2 - \alpha_j^2$ and $s_{ij} = \alpha_i - \alpha_j$. This procedure may be repeated again with the three new selected points. Under some regularity hypothesis, the convergence rate of this method is super-linear [2].

Another approach consists in differentiating the cost function towards the direction of descent with a Taylor expansion, andin neglecting second order derivatives:

$$g'(\alpha) = \frac{d}{d\alpha}\left\|u\left(\psi^k + \alpha d^k\right) - u_d\right\|_{\mathcal{X}}^2 = \frac{d}{d\alpha}\left\|u\left(\psi^k\right) - \alpha u'\left(\psi^k; d^k\right) - u_d\right\|_{\mathcal{X}}^2 = 0 \tag{12}$$

with $u'\left(\psi, d^k\right) = u'$ the derivative of u at the point ψ^k and in the direction d^k. This equation gives straightforwardly:

$$\left(u', u - u_d\right)_{\mathcal{X}} + \alpha\left(u', u'\right)_{\mathcal{X}} = 0 \tag{13}$$

$$\alpha = -\frac{(u', u - u_d)_{\mathcal{X}}}{(u', u')_{\mathcal{X}}} \tag{14}$$

This latter method – which is widely used in the heat transfer community – can give easily an accurate step size α when the cost function j is close to quadratic, *i.e.* when the state u varies almost linearly with ψ.

4.5 Other methods – Inexact line-search

A great number of other one-dimensional optimization methods may be found in the literature. These methods may be more or less complicated and some of them may be much more optimal than the above-presented methods. In practice the Fibonacci method,the golden section search method and the cubic interpolation method are also very widely used in practice (the reader may refer to [4, 2] for more details). All these methods can be quite CPU-time consuming, and in fact, the convergence of some of the methods presented afterwards in Section 5 (typically the BFGS method) can be reached without getting a point very close to satisfying $g(\alpha) = 0$. Well-accepted conditions used to build inexact line-search algorithms are based on the two rules:

a) α must not be too large in order, for instance, to avoid oscillations,

b) α must not be chosen too small in order to prevent from premature convergence.

Among the large number of inexact line-search algorithms, one is based on the Goldstein rules (see Figure 7) which first ensures condition a) by satisfying (15) choosing $m_1 \in [0, 1]$, and second ensures condition b) satisfying (16) choosing $m_2 \in [m_1, 1]$.

$$g(\alpha) \leq g(0) + m_1 \alpha g'(0) \tag{15}$$

$$g(\alpha) \geq g(0) + m_2 \alpha g'(0) \tag{16}$$

Other rules can be stated in similar ways. For instance, the Armijo's method is a variant of the Golstein method. Related algorithms are very simple and can be found in any book on optimization.

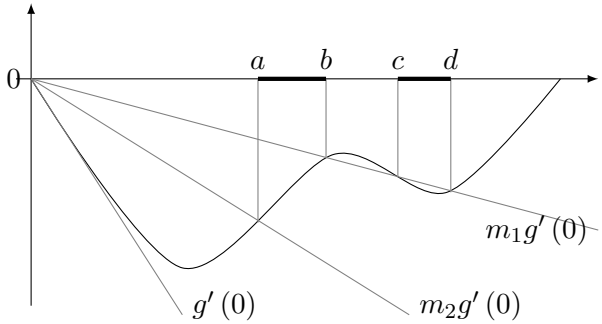


Figure 7: Set of points satisfying the Goldstein's rules: $[a, b] \cup [c, d]$.

Algorithm 1: Typical method based on the Goldstein rules

input : $\alpha_{\min} = 0,\ \alpha_{\max} = \infty, \psi = \psi^k, \nabla j, d = d^k$
output: $\bar{\alpha}$

1 Give some initial value to α;
Compute $g'(0) = (\nabla j, d)$;
2 Compute $g(\alpha) = j(\psi + \alpha d)$;
if $g(\alpha) \leq g(0) + m_1 \alpha g'(0)$ **then**
| go to 3)
else
| set $\alpha_{\max} = \alpha$ and go to 5
end
3 Compare $g(\alpha)$ and $g(0) + m_2 \alpha g'(0)$;
if $g(\alpha) \geq g(0) + m_2 \alpha g'(0)$ **then**
| END
else
| go to 4
end
4 Set $\alpha_{\min} = \alpha$;
5 Look for new value in $]\alpha_{\min}, \alpha_{\max}[$ and return to 2

5 Gradient-type n-dimensional optimization algorithms

Since in all cases, the stationarity of j is a necessary optimality condition, almost all unconstrained optimization methods consist in searching the stationary point $\bar{\psi}$ where $\nabla j(\bar{\psi}) = 0$. The usual methods are iterative and proceed this way: one generates a sequence of points ψ^0, ψ^1,... ψ^p which converges to a local optimum of j. At each stage p, ψ^{p+1} is defined by $\psi^{p+1} = \psi^p + \alpha^p d^p$ where d^p is a displacement direction which may be either the opposite of the gradient of j at ψ^p (i.e. $d^p = -\nabla j(\psi^p)$), or computed from the gradient, or chosen in any another way, provided that it is a descent direction, i.e. satisfying $(\nabla j(\psi^p), d^p) < 0$.

5.1 1st order gradient methods

5.1.1 The gradient with predefined steps method (1st order method)

At each iteration step p, the gradient $\nabla j(\psi^p)$ gives the direction of the largest increase of j. The procedure consists in computing the gradient, and in finding the new point according to the predefined strictly positive step size α^p as:

$$\psi^{p+1} = \psi^p - \alpha^p \frac{\nabla j(\psi^p)}{\|\nabla j(\psi^p)\|}. \tag{17}$$

It may be shown that this iterative scheme converges to $\bar{\psi}$ provided that $\alpha^p \to 0 \quad (p \to \infty)$ and $\sum_{p=0}^{\infty} \alpha^p = +\infty$. One can choose for instance $\alpha^p = 1/p$. The main drawback of this method is its very low convergence rate.

5.1.2 The steepest descent method (1st order method)

In this frequently used method, α^p is chosen at each iteration p so as to minimize the function $g(\alpha) = j\left(\psi^p - \alpha \nabla j(\psi^p)\right)$ on the set of $\alpha \geq 0$. The algorithm is thus the following. One chooses a starting point ψ^0 and set $p = 0$. At each iteration p, one computes the gradient and set $d^p = -\nabla j(\psi^p)$. One then solves the one-dimensional problem (see section 4) and set $\psi^{p+1} = \psi^p + \alpha^p d^p$. This procedure is repeated until a stopping test is satisfied (see section 2.5). The main disadvantage of the steepest descent method is the fact

that the convergence can still be very slow. As a matter of fact, since α^p minimizes $g(\alpha) = j\left(\psi^p + \alpha d^p\right)$ then $g'(\alpha^p) = \left(d^p, \nabla j\left(\psi^p + \alpha d^p\right)\right) = \left(d^p, \nabla j\left(\psi^{p+1}\right)\right)$. Hence $\left(d^p, d^{p+1}\right) = 0$. This means that two successive displacements are strictly orthogonal. As a direct consequence, the number of steps to minimize elongated valley-type functions for instance may be very high (see fig. 8 and then fig. 10d on page 22).

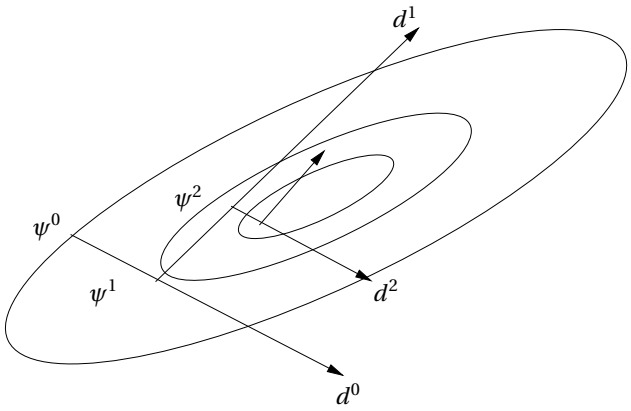


Figure 8: When the steepest descent method is used, the two consecutive directions are orthogonal.

5.1.3 The conjugate gradient method for quadratic functions (1st order method)

In this section we shall firstly assume that the cost function is quadratic. The case of arbitrary functions shall be dealt with in section 5.1.4. Let the quadratic functional be of the form:

$$j(\psi) = \frac{1}{2}\left(\mathscr{A}\psi, \psi\right), \tag{18}$$

and let us recall the definition for two conjugate vectors. Let $\mathscr{A}$ be a given symmetric matrix (operator). Two vectors x_1 and x_2 are $\mathscr{A}$-conjugate if $(\mathscr{A}x_1, x_2) = 0$. The general method to optimize j is the following. Let us start with a given ψ^0 and choose $d^0 = -\nabla j(\psi^0)$. One may remark that for quadratic functions, the one-dimensional minimization procedure may be analytically solved. Recalling that the minimization of $g(\alpha)$ along the direction d^0 should lead to the fact that this current direction (d^0) would be orthogonal to the next gradient $\nabla j(\psi^1)$, one has:

$$\left(d^0, \nabla j\left(\psi^1\right)\right) = 0. \tag{19}$$

Using the relationship $\nabla j(\psi) = \mathscr{A}\psi$ given by the differentiation of (18) and the reactualization formulation $\psi^1 = \psi^0 + \alpha^0 d^0$, (19) becomes:

$$\begin{aligned} \left(d^0, \nabla j\left(\psi^1\right)\right) &= \left(d^0, \mathscr{A}\psi^1\right) \\ &= \left(d^0, \mathscr{A}\left(\psi^0 + \alpha^0 d^0\right)\right) \\ &= \left(d^0, \mathscr{A}\psi^0\right) + \alpha^0\left(d^0, \mathscr{A}d^0\right). \end{aligned} \tag{20}$$

Equaling (20) to zero gives the step size α^0:

$$\alpha^0 = -\frac{\left(d^0, \mathscr{A}\psi^0\right)}{\left(d^0, \mathscr{A}d^0\right)}. \tag{21}$$

Next, at stage p, we are at the point ψ^p and we compute the gradient $\nabla j(\psi^p)$. The direction d^p is obtained by combining linearly the gradient $\nabla j(\psi^p)$ and the previous direction d^{p-1}, where the coefficient β^p is chosen in such a way that d^p is $\mathscr{A}$-conjugate to the previous direction. Hence:

$$\begin{aligned} \left(d^p, \mathscr{A}d^{p-1}\right) &= \left(-\nabla j\left(\psi^p\right) + \beta^p d^{p-1}, \mathscr{A}d^{p-1}\right) \\ &= -\left(\nabla j\left(\psi^p\right), \mathscr{A}d^{p-1}\right) + \beta^p\left(d^{p-1}, \mathscr{A}d^{p-1}\right). \end{aligned} \tag{22}$$

Next, choosing β^p such that the previous equation equals zero yields to:

$$\beta^p = \frac{\left(\nabla j(\psi^p), \mathscr{A} d^{p-1}\right)}{\left(d^{p-1}, \mathscr{A} d^{p-1}\right)}. \tag{23}$$

The algorithm based on the above relationships is given in algorithm 2. Also, it is proved, see [2], that the conjugate gradient method applied to quadratic functions converges in at most n iterations, where $n = \dim \psi$.

Algorithm 2: The conjugate gradient algorithm applied to quadratic functions

1. Let $p = 0$, ψ^0 be the starting point,
   compute the gradient and the descent direction, $d^0 = -\nabla j(\psi^0)$,
   compute the step size $\alpha^0 = -\dfrac{\left(d^0, \mathscr{A}\psi^0\right)}{\left(d^0, \mathscr{A} d^0\right)}$;
2. At step p, we are at the point ψ^p.
   We define $\psi^{p+1} = \psi^p + \alpha^p d^p$ with:
   - the step size $\alpha^p = -\dfrac{(d^p, \nabla j(\psi^p))}{(d^p, \mathscr{A} d^p)}$
   - the direction $d^p = -\nabla j(\psi^p) + \beta^p d^{p-1}$
   - where the coefficient needed for conjugate directions: $\beta^p = \dfrac{\left(\nabla j(\psi^p), \mathscr{A} d^{p-1}\right)}{\left(d^{p-1}, \mathscr{A} d^{p-1}\right)}$;
3. Stopping rule (see Section 2.5). If satisfied: End, otherwise set $p \leftarrow p + 1$ and return to step (2).

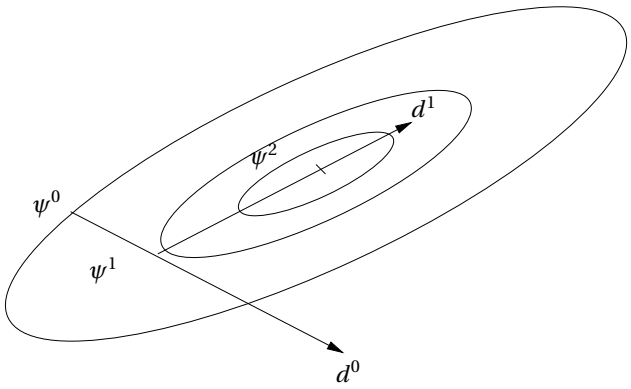


Figure 9: When the conjugate gradient method is used, the two consecutive directions are conjugate instead of orthogonal. Applied to a quadratic function, the method converges in at most n iterations (in this figure two iterations are needed since $\dim \psi = 2$).

5.1.4 The conjugate gradient method for arbitrary (non quadratic) functions (1st order)

Before presenting the application of conjugate gradient methods on arbitrary functions, let us give some properties inherent to quadratic functions. Differentiating eq. (18), and taking into account of the re-

actualization relationship, one has:

$$\begin{aligned}\nabla j(\psi^p) - \nabla j(\psi^{p-1}) &= \mathscr{A}\left(\psi^p - \psi^{p-1}\right)\\ &= \mathscr{A}\left(\psi^{p-1} + \alpha^{p-1}d^{p-1} - \psi^{p-1}\right)\\ &= \alpha^{p-1}\mathscr{A}d^{p-1},\end{aligned} \tag{24}$$

which also gives the following relationship:

$$\frac{1}{\alpha^{p-1}}\left(\nabla j(\psi^p), \nabla j(\psi^p) - \nabla j(\psi^{p-1}\right) = \left(\nabla j(\psi^p), \mathscr{A}d^{p-1}\right). \tag{25}$$

On the other hand, substituting (25) into (23) gives

$$\beta^p = \frac{\left(\nabla j(\psi^p), \mathscr{A}d^{p-1}\right)}{(d^{p-1}, \mathscr{A}d^{p-1})} = \frac{\left(\nabla j(\psi^p), \nabla j(\psi^p) - \nabla j(\psi^{p-1})\right)}{(d^{p-1}, \nabla j(\psi^p) - \nabla j(\psi^p - 1))}. \tag{26}$$

Next, expanding the descent direction d^{p-1}, (26) becomes:

$$\beta^p = \frac{\left(\nabla j(\psi^p), \nabla j(\psi^p) - \nabla j(\psi^{p-1})\right)}{(-\nabla j(\psi^{p-1}) + \beta^{p-1}d^{p-2}, \nabla j(\psi^p) - \nabla j(\psi^{p-1}))}; \tag{27}$$

$$\beta^p = \frac{\left(\nabla j(\psi^p), \nabla j(\psi^p) - \nabla j(\psi^{p-1})\right)}{(-\nabla j(\psi^{p-1}) - \beta^{p-1}\nabla j(\psi^{p-2}) + \Lambda, \nabla j(\psi^p) - \nabla j(\psi^{p-1}))}, \tag{28}$$

where Λ is the series given from the re-actualizations. All the gradients being orthogonal to each other, (28) becomes:

$$\beta^p = \frac{\left(\nabla j(\psi^p), \nabla j(\psi^p) - \nabla j(\psi^{p-1})\right)}{(\nabla j(\psi^{p-1}), \nabla j(\psi^{p-1}))}, \tag{29}$$

and also:

$$\beta^p = \frac{(\nabla j(\psi^p), \nabla j(\psi^p))}{(\nabla j(\psi^{p-1}), \nabla j(\psi^{p-1}))}. \tag{30}$$

It is pointed out that in the neighborhood of the optimum, non-quadratic functions may always be approximated by quadratic functions. The Fletcher and Reeves' method consists in applying (30) to access β^p while the Polak and Ribiere's method consists in applying (29) to access β^p. Taking into account of above remarks, the *conjugate gradient algorithm applied to arbitrary functions* is given in Algorithm 3. It is important to note that the global convergence of the presented methods is only ensured if a periodic restart is carried out . The restart $d^n = -\nabla j(u^n)$ is usually carried out every n iterations, at least.

Algorithm 3: The conjugate gradient algorithm applied to arbitrary functions

1. Let $p = 0$, ψ^0 be the starting point, $d^0 = -\nabla j(\psi^0)$;
2. At step p, we are at the point ψ^p; we define $\psi^{p+1} = \psi^p + \alpha^p d^p$ with:
   - the step size $\alpha^p = \arg\min_{\alpha\in\mathbb{R}^+} g(\alpha) = j\left(\psi^p + \alpha d^p\right)$ with:
   - the direction $d^p = -\nabla j(\psi^p) + \beta^p d^{p-1}$ where
   - the conjugate condition β^p satisfies either (29) or (30) depending on the chosen method;
3. Stopping rule (see subsection 2.5). If satisfies: End, otherwise, set $p \leftarrow p + 1$ and return to step (2).

5.2 The Newton's method (2nd order)

Let us assume that the cost function $j(\psi)$ is now twice continuously differentiable and that second derivatives exist. The idea is to approach the next cost function gradient by its quadratic approximation through a Taylor development:

$$\nabla j(\psi^{p+1}) = \nabla j(\psi^p) + \left[\nabla^2 j(\psi^p)\right] \delta\psi^p + \mathcal{O}\left(\delta\psi^p\right)^2, \tag{31}$$

and equaling the obtained approximated gradient to zero to get the new parameter $\psi^{p+1} = \delta\psi^p + \psi^p$:

$$\psi^{p+1} = \psi^p - \left[\nabla^2 j\left(\psi^p\right)\right]^{-1} \nabla j(\psi^p). \tag{32}$$

Note that while using second-order optimization algorithms, the direction of descent as well as the step size are obtained from (32) in one go. Another interesting point is the fact that the algorithm converges to $\bar{\psi}$ in a single step when applied to strictly quadratic functions. However, for arbitrary functions, $\mathcal{O}\left(\delta\psi^p\right)^2$ may be far from zero in eq. (31); yielding to some errors in the displacement $\delta\psi^p$, and thus in the new point ψ^{p+1}. As a consequence, if the starting point ψ^0 is too far away from the solution $\bar{\psi}$, then the Newton method may not converge. On the other hand, since the approximation of $j(\psi)$ by a quadratic function is almost always valid in the neighborhood of $\bar{\psi}$, then the algorithm should converge to $\bar{\psi}$ if the starting point ψ^0 is chosen closely enough to the solution. Moreover, it is very common to control the step size this way. One first calculates the direction $d^p = -\left[\nabla^2 j(\psi^p)\right]^{-1} \nabla j(\psi^p)$ and control the step size through an iterative one-dimensional minimization problem of the kind $\min g(\alpha) = j\left(\psi^p + \alpha d^p\right)$ before the actualization $\psi^{p+1} = \psi^p + \alpha d^p$. One limitation of the Newton's method is when the Hessian $\nabla^2 j(u^p)$ is not positive definite. In these cases, the direction given by $d^p = -\left[\nabla^2 j(\psi^p)\right]^{-1} \nabla j(u^p)$ may not be a descent direction, and the global convergence of the algorithm may not be guaranteed any more. Moreover, and above all, the Hessian is usually very difficult to compute and highly time consuming. To overcome these difficulties, one should prefer using one of the numerous quasi-Newton methods detailed afterwards.

5.3 Quasi-Newton methods

Quasi-Newton methods consist in generalizing the Newton's recurrence formulation (32). Since the limitation of the Newton's method is the restriction of the Hessian $\nabla^2 j(u^p)$ to be positive definite, the natural extension consists in replacing the *inverse of the Hessian* by an approximation to a positive definite matrix denoted $\mathbf{H}^p$. Obviously, this matrix is modified at each step p. There is much flexibility in the choice for computing the matrix $\mathbf{H}^p$. In general, the condition given by (33) is imposed:

$$\mathbf{H}\left[\nabla j(\psi^p) - \nabla j(\psi^{p-1})\right] = \psi^p - \psi^{p-1}. \tag{33}$$

Various corrections of the type

$$\mathbf{H}^{p+1} = \mathbf{H}^p + \Lambda^p \tag{34}$$

may be found in the literature [2]. Depending on whether Δ^p is of rank 1 or 2, we shall speak of a correction of rank 1 or 2.

5.3.1 Rank 1 correction

The point is to choose a symmetric matrix $\mathbf{H}^0$ and to perform the corrections so that they preserve the symmetry of the matrices $\mathbf{H}^p$. The rank 1 correction matrix consists in choosing $\Delta^p = \alpha^p v^p v^{p\top}$ where v^p is a vector and α^p is a scalar such that, from a symetric matrix $\mathbf{H}^0$, the correction preserves the symetry of matrices $\mathbf{H}^p$. Denoting

$$\delta^p = \psi^{p+1} - \psi^p \tag{35}$$

$$\gamma^p = \nabla j(\psi^{p+1}) - \nabla j(\psi^p) \tag{36}$$

one chooses α^p and v^p such that $\mathbf{H}^{p+1}\gamma^p = \delta^p$, thus:

$$\left[\mathbf{H}^p + \alpha^p(v^p v^{pt})\right]\gamma^p = \delta^p, \tag{37}$$

and

$$\gamma^{p\top}\mathbf{H}^p\gamma^p + \alpha^p\left(\gamma^{p\top}v^p\right)\left(v^{p\top}\gamma^p\right) = \gamma^{p\top}\delta^p, \tag{38}$$

thus

$$\alpha^p\left(v^{p\top}\gamma^p\right)^2 = \gamma^{p\top}\left(\delta^p - \mathbf{H}^p\gamma^p\right). \tag{39}$$

Using the identity

$$\alpha^p\left(v^p v^{p\top}\right) = \frac{\left(\alpha^p v^p v^{p\top}\gamma^p\right)\left(\alpha^p v^p v^{p\top}\gamma^p\right)^\top}{\alpha^p\left(v^{p\top}\gamma^p\right)^2}, \tag{40}$$

and using (37) and (38) to get

$$\alpha^p v^p v^{p\top}\gamma^p = \delta^p - \mathbf{H}^p\gamma^p, \tag{41}$$

$$\alpha^p\left(v^{p\top}\gamma^p\right)^2 = \gamma^{p\top}\left(\delta^p - \mathbf{H}^p\gamma^p\right), \tag{42}$$

one obtains the correction (of rank 1) of the inverse Hessian:

$$\mathbf{H}^{p+1} - \mathbf{H}^p = \alpha^p\left(v^p v^{p\top}\right) = \frac{\left(\delta^p - \mathbf{H}^p\gamma^p\right)\left(\delta^p - \mathbf{H}^p\gamma^p\right)^\top}{\gamma^{p\top}\left(\delta^p - \mathbf{H}^p\gamma^p\right)}. \tag{43}$$

5.3.2 The rank 2 Davidon-Fletcher-Powell (DFP) algorithm

The Davidon-Fletcher-Powell algorithm (in short DFP) consists in modifying the inverse Hessian with the correction formulation of rank 2:

$$\mathbf{H}^{p+1} = \mathbf{H}^p + \frac{\delta^p(\delta^p)^\top}{(\delta^p)^\top\gamma^p} - \frac{\mathbf{H}^p\gamma^p(\gamma^p)^\top\mathbf{H}^p}{(\gamma^p)^\top\mathbf{H}\gamma^p} \tag{44}$$

where we have defined above $\delta^p = \psi^{p+1} - \psi^p$ and $\gamma^p = \nabla j(\psi^{p+1}) - \nabla j(\psi^p)$, and where the new point ψ^{p+1} is obtained from ψ^p through the displacement

$$d^p = -\mathbf{H}^p\nabla j(\psi^p). \tag{45}$$

The global DFP method is presented in algorithm 4.

5.3.3 The rank 2 Broyden – Fletcher – Goldfarb – Shanno (BFGS) algorithm

The Broyden – Fletcher – Goldfarb – Shanno algorithm (in short BFGS) developped in 1969-1970 uses a rank 2 correction matrix for the inverse Hessian that is derived from eq. (44). It can be shown [2] that the vectors δ^p and γ^p can permute in eq. (44) and in the relationship $\mathbf{H}^{p+1}\gamma^p = \delta^p$. The correction eq. (44) can thus also approximate the Hessian itself, and the correction for the inverse Hessian $\mathbf{H}^{p+1}$ can thus be given from $\mathbf{H}^p$ through the correction formulation:

$$\mathbf{H}^{p+1} = \mathbf{H}^p + \left[1 + \frac{\gamma^{pt}\mathbf{H}^p\gamma^p}{\delta^{pt}\gamma^p}\right]\frac{\delta^p(\delta^p)^t}{(\delta^p)^t\gamma^p} - \frac{\delta^p\gamma^{pt}\mathbf{H}^p + \mathbf{H}^p\gamma^p\delta^{pt}}{\delta^{pt}\gamma^p}. \tag{47}$$

When applied to a non purely quadratic function, one has, as for the conjugate gradient method and the DFP method, to carry out a periodic restart in order to ensure the convergence [4, 11]. It is known that the BFGS algorithm is superior than the DFP algorithm is the sense that it is much less sensitive on the line-search inaccuracy, allowing the use of economical inexact line-search algorithms [2].

Algorithm 4: The Davidon – Fletcher – Powell (DFP) algorithm

1. Let $p = 0$, ψ^0 be the starting point. Choose any positive definite matrix $\mathbf{H}^0$ (often the identity matrix);
2. at step p, compute the displacement direction $d^p = -\mathbf{H}^p \nabla j(\psi^p)$, and find ψ^{p+1} at the minimum of $j(\psi^p + \alpha d^p)$ with $\alpha \geq 0$;
3. set $\delta^p = \psi^{p+1} - \psi^p$ and compute $\gamma^p = \nabla j(\psi^{p+1}) - \nabla j(\psi^p)$ to actualize:

$$\mathbf{H}^{p+1} = \mathbf{H}^p + \frac{\delta^p (\delta^p)^t}{(\delta^p)^t \gamma^p} - \frac{\mathbf{H}^p \gamma^p (\gamma^p)^t \mathbf{H}^p}{(\gamma^p)^t \mathbf{H} \gamma^p}; \tag{46}$$

4. Stopping rule (see section 3.4). If satisfies: End, otherwise, set $p \leftarrow p + 1$ and return to step item 2.

Algorithm 5: The BFGS algorithm

1. Let $p = 0$, ψ^0 be the starting point. Choose any positive definite matrix $\mathbf{H}^0$ (often the identity matrix);
2. at step p, compute the displacement direction $d^p = -\mathbf{H}^p \nabla j(\psi^p)$, and find ψ^{p+1} at the minimum of $j(\psi^p + \alpha d^p)$ with $\alpha \geq 0$;
3. set $\delta^p = \psi^{p+1} - \psi^p$ and compute $\gamma^p = \nabla j(\psi^{p+1}) - \nabla j(\psi^p)$ to actualize:

$$\mathbf{H}^{p+1} = \mathbf{H}^p + \left[1 + \frac{\gamma^{pt} \mathbf{H}^p \gamma^p}{\delta^{pt} \gamma^p}\right] \frac{\delta^p (\delta^p)^t}{(\delta^p)^t \gamma^p} - \frac{\delta^p \gamma^{pt} \mathbf{H}^p + \mathbf{H}^p \gamma^p \delta^{pt}}{\delta^{pt} \gamma^p} \tag{48}$$

4. Stopping rule (see section 3.4). If satisfies: End, otherwise, set $p \leftarrow p + 1$ and return to step item 2.

5.3.4 Gauss–Newton

When the cost function is explicitly a square norm of the error between the prediction and the state, that is of the form

$$j(\psi) := \mathcal{J}(u) = \|u - u_d\|^2_{\mathcal{X}}\,, \tag{49}$$

then the Gauss–Newton method or some derivatives or it (e.g. Levenberg–Marquardt) may be interesting to deal with, especially if the number of parameters is small. Before going deeper into the cost function gradient computation (see section 6), defining $u'(\psi; \delta\psi)$ as the derivative of the state at the point ψ in the direction $\delta\psi$ as:

$$u'(\psi; \delta\psi) := \lim_{\epsilon \to 0} \frac{u(\psi + \epsilon\delta\psi) - u(\psi)}{\epsilon}, \tag{50}$$

then the directional derivative of the cost function writes out as:

$$j'(\psi; \delta\psi) = \big(u - u_d, u'(\psi; \delta\psi)\big)_{\mathcal{X}}\,, \tag{51}$$

where $j'(\psi; \delta\psi) = (\nabla j(\psi), \delta\psi)$. In the analogue way, the second derivative of $j(\psi)$ at the point ψ in the directions $\delta\psi$ and $\delta\phi$ is given by:

$$j''(\psi; \delta\psi, \delta\phi) = \big(u - u_d, u''(\psi; \delta\psi, \delta\phi)\big)_{\mathcal{X}} + \big(u'(\psi; \delta\psi), u'(\psi; \delta\phi)\big)_{\mathcal{X}}\,. \tag{52}$$

Neglecting the second-order term (this is actually the Gauss–Newton approach), we have:

$$j''(\psi;\delta\psi,\delta\phi) \approx \left(u'(\psi;\delta\psi), u'(\psi;\delta\phi)\right)_{\mathcal{X}}. \tag{53}$$

In order to build up the gradient vector of cost function and the approximated Hessian matrix, one has to choose the directions for the whole canonical basis of ψ. Doing so, one can use the so-called sensititivity matrix S which gathers the derivatives of u in all directions $\delta\psi_i$, $i = 1, \ldots, \dim\psi$, and the product $(u'(\psi;\delta\psi_i), u'(\psi;\delta\psi_j))$ involved in (53) is the product of the so-called sensitivity matrix with its transposed. The Newton relationship is thus approximated as:

$$S^t S\delta\psi^k = -\nabla j(\psi^k). \tag{54}$$

The matrix system S^tS is obviously symmetric and positive definite with a dominant diagonal yielding thus to interesting features (Cholesky factorization, etc.). Though the Gauss–Newton system eq. (54) presents inherent interesting features (it almost gives in one step the descent direction and the step size), the matrix S^tS is likely to be ill-conditionned. One way to decrease significantly the ill-condition feature is to "damp" the system, using:

$$\left[S^tS + \ell I\right]\delta\psi^k = -\nabla j(\psi^k), \tag{55}$$

or better:

$$\left[S^tS + \ell\mathrm{diag}(S^tS)\right]\delta\psi^k = -\nabla j(\psi^k). \tag{56}$$

Note that $\ell \to 0$ yields the Gauss–Newton algorithm while ℓ bigger gives an approximation of the steepest descent gradient algorithm. In practice, the parameter ℓ may be adjusted at each iteration.

5.4 Elements of comparison between some presented methods

Some of the presented methods are below tested on the well-known Rosenbrock function:

$$f(x,y) = (x-\alpha)^2 + \beta(x^2-y)^2. \tag{57}$$

For the considered case, the chosen parameters are $\alpha = 1$ and $\beta = 100$, so that the optimum is at $(1,1)$. Figure 10a on page 22 presents the function. This function presents a long elongated valley where the function gradient is very low. Next, the PSO algorithm is the one from [10].

The deterministic simplex method from the GSL library starting from the point $x^0 = -1$, $y^0 = 1$ needs 64 evaluations of the cost function. The stopping criterion is based on the simplex characteristic size equal to 10^{-2}. The PSO algorithm taken from [10] with 20 particles with 3 informed particles, $\phi = 4.14$, $\chi = \frac{2}{\phi-2+\sqrt{\phi^2-4\phi}}$, $\lambda_1 = \lambda_2 = 0.5\chi\phi$. The stopping criterion is based on the cost function equal to 10^{-5}. With these parameters, around 6,000 evaluations of the cost function is needed for the minimization. For the Steepest descent, the conjugate gradient and the BFGS algorithms, the stopping criterion is based either on the gradient norm equal to 10^{-3}, or on a maximum number of iterations equal to 10,000. For the steepest descent method, the maximum of iteration criterion is achieved. For the conjugate gradient, and the BFGS method, 49 and 11 iterations are needed, respectively.

To sum up about this numerical optimization test case in which we were searching the minimum of the Rosenbrock function, we can give the following comments and conclusions:

- The PSO method, which is a stochastic zero-order method – as genetic algorithms are also – does converge to the minimum, but at a huge expense. In fact, usually, such stochastic methods are even able to find the minimum of non-convex functions, which is their most important advantage, but anyway at the price of being very expensive.

- When the function is likely to be convex (which is not the case of the Rosenbrock function), one should prefer less expensive deterministic optimization algorithms. Among those, the simplex method is also a gradient-free (zero-order optimizer) so it finds the minimum of the convex function at a more moderate expense, because it is deterministic. But we are here – in this simple example – handling a function of only two parameters, which is very few, and tens of function evaluations are necessary. With more parameters, say hundreds or thousands (this is at least what one usually has in function estimation), zero-order gradient-free are still too expensive and thus cannot be used in practice; gradient-based optimizers should be prefered.
- When the function is likely to be convex, and when the cost function depends on some states – solution of partial differential equations –, then the model itself is likely to be differentiated. In such cases, gradient-based optimizers are to be chosen. Among those, with respect to the most basic steepest descent algorithm, the numerical effort of implementing the conjugate gradient, or better the BFGS, is highly recommended.
- The example presented here, on the only two-dimensional Rosenbrock function, has demonstrate this result. Such conclusions are of course much solid when it comes to function estimation where higher dimensions are encountered.

6 Cost function gradient

We recall here that the function to be minimized is the cost function $\mathcal{J}(u)$, expressed in terms of the state u, but minimized with respect to the parameters ψ. We thus have the equality (by definition) between the cost function and its reduced version: $j(\psi) := \mathcal{J}(u)$. The state u is related to the parameters ψ through an operator (which may be linear, or not) that combines the partial differential equations along with the boundary conditions, initial conditions, etc. This operator is denoted as $\mathcal{S}$ for the state problem. To be concise, one writes down

$$\mathcal{S}(u, \psi) = 0, \tag{58}$$

where we have the mapping $\psi \mapsto u(\psi)$. Often, the space (and time) is discretized so that the state operator $\mathcal{S}$ is approximated in some matrix formulation. In this case, we have $\mathcal{R}(u, \psi) = 0$, with $\dim \mathcal{R} = \dim u$. Note that u involved in (58) is continuous while u involved in $\mathcal{R}(u, \psi) = 0$ is likely to be already discretized (using finite difference, finite elements, etc.). We now need the definition of the directional derivative of $j(\psi)$ in the direction $\delta\psi$ (see definition 6). Other kinds of derivatives can also be used, such as the Gâteaux or Fréchet derivatives, see [1] for technical definitions.

Definition 6 (Directional derivative)**.** *Let a point $\psi \in \mathcal{K}$ and a direction $\phi \in \mathcal{K}$. One defines $\ell(t) := \psi + t\phi$ and the function $\mathscr{J}(t) := j(\ell(t))$. The directional derivative of j at the point ψ in the direction ϕ is:*

$$j'(\psi; \phi) := \mathscr{J}'(0) = \lim_{\substack{t \to 0 \\ t > 0}} \frac{j(\psi + t\phi) - j(\psi)}{t}. \tag{59}$$

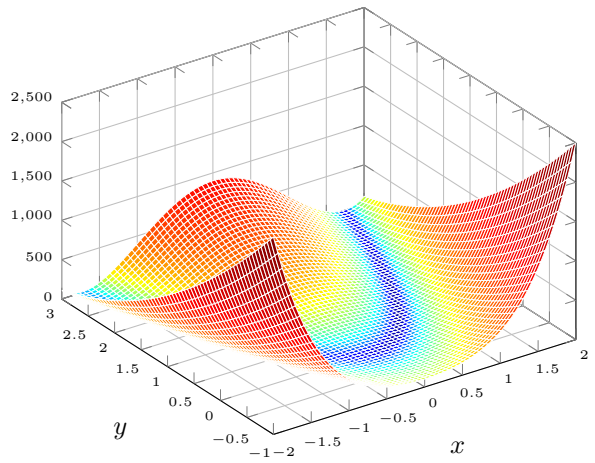
It has been seen before (see eq. (51)) that we have the equality

$$j'(\psi; \delta\psi) = \big(u - u_d, u'(\psi; \delta\psi)\big)_{\mathcal{X}}, \tag{60}$$

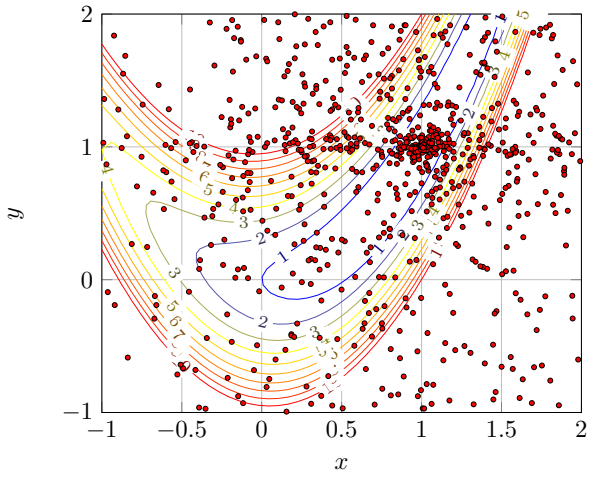
and, because of linearity of both $u'(\psi; \delta\psi)$ and $j'(\psi; \delta\psi)$ in $\delta\psi$:

$$j'(\psi; \delta\psi) = (\nabla j(\psi), \delta\psi)_{\mathcal{Z}}. \tag{61}$$

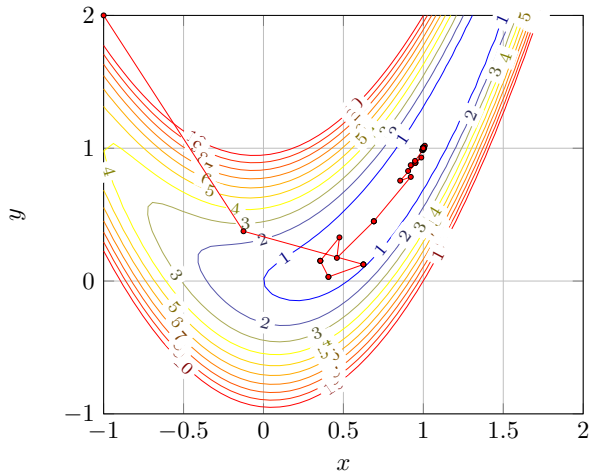
where $\mathcal{Z}$ is most of the time chosen equal to $\mathcal{Y}$ but it can be chosen differently for regularization purposes.



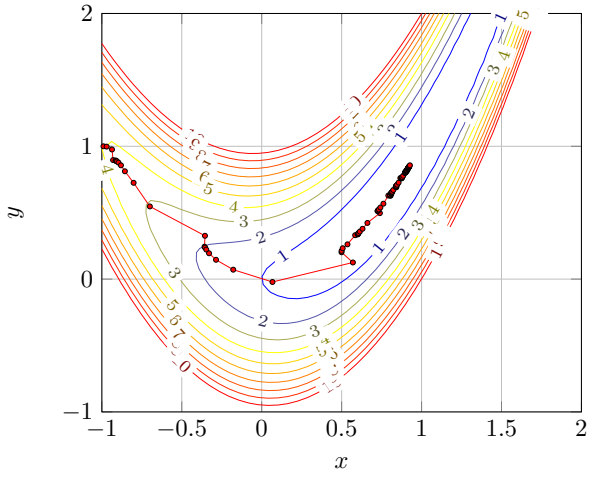
(a) 2–D Rosenbrock function.



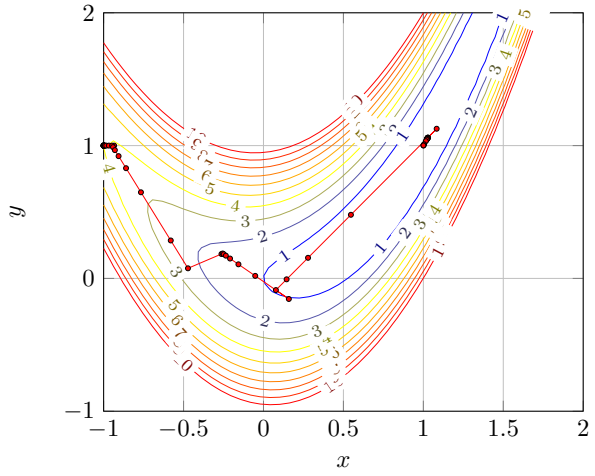
(b) PSO algorithm: ≈ 6,000 cost function evaluations.



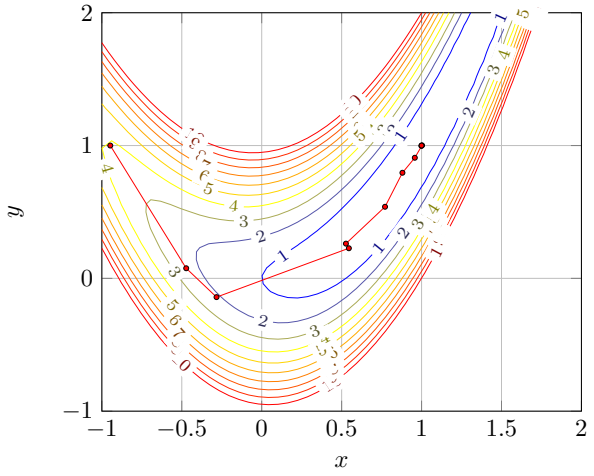
(c) Simplex algorithm: 64 cost function evaluations.



(d) Steepest descent algorithm: more than 100 cost function evaluations.



(e) Conjugate gradients descent algorithm: 45 cost function evaluations.



(f) BFGS descent algorithm: 11 cost function evaluations.

Figure 10: Numerical comparison of optimizers on the 2-D Rosenbrock function.

6.1 Finite difference

The finite difference approach consists in approaching the cost function gradient through a substraction of the cost function with a perturbed cost function for the whole canonical base of ψ, that is $\delta\psi = \delta\psi_1, \delta\psi_2, \ldots, \delta\psi_{\dim\psi}$. For the i^{th} component, we have:

$$(\nabla j(\psi))_i = (\nabla j(\psi), \delta\psi_i)_{\mathcal{Z}} \approx \frac{j(\psi + \epsilon\delta\psi_i) - j(\psi)}{\epsilon}. \tag{62}$$

Usually, in order to perform the same relative perturbation on all components ψ_i, one rather uses $\epsilon_i \leftarrow \varepsilon|\psi_i|$, where the positive scalar ε is fixed. The very simple related algorithm is described in algorithm 6.

Algorithm 6: The finite difference algorithm to compute the gradient of the cost function

Set the length $\varepsilon > 0$;
At iteration p, compute the state $u(\psi^p)$, compute $j(\psi^p)$;
foreach $i = 1, \ldots, \dim\psi$ **do**
  Compute the cost $j(\psi^p + \varepsilon|\psi_i|\delta\psi_i)$;
  Set the gradient $(\nabla j(\psi))_i \leftarrow \dfrac{j(\psi^p + \varepsilon|\psi_i|\delta\psi_i) - j(\psi^p)}{\varepsilon|\psi_i|}$
end
Integrate the gradient within the optimization methods that do not rely on the sensitivities (conjugate gradient or BFGS for instance among the presented methods)

In practice, the tuning parameter ε has to be chosen within a region where variables depend roughly linearly on ε. Indeed for too small values, the round-off errors dominate while for too high values one gets a nonlinear behavior. Even though the finite difference method is easy to implement, it has the disadvantage of being highly CPU time consuming. Indeed, the method needs as many integrations of the model $\mathcal{S}(u, \psi) = 0$ as the number of parameters, $\dim\psi$. The gradient computed this way can be integrated to the previously presented optimization methods that do not rely on u', such as the conjugate gradient methods, or better the BFGS.

When performing the finite differenciation with respect to ψ_i, one also accesses the approximated perturbed state $u'(\psi; \delta\psi_i)$. This way, one can use again the conjugate gradient methods or the BFGS method for instance, but also the Gauss–Newton-type methods based on matrix inversion and which do rely on the sensitivities $u'(\psi; \delta\psi_i)$, $i = 1, \ldots, \dim\psi$. Doing so, the related optimization is given in algorithm 7.

6.2 Forward differentiation

The forward differentiation approach consists in computing $u'(\psi; \delta\psi_i)$ differentiating the state equations $\mathcal{S}(u, \psi) = 0$, to get:.

$$\mathcal{S}'_u(u, \psi)u' + \mathcal{S}'_\psi(u, \psi)\delta\psi = 0. \tag{63}$$

As in the previous section, the gradient computation needs one integration of eq. (63) per parameter ψ_i; so one needs $\dim\psi$ integrations in total to access the full gradient $\nabla j(\psi)$. However, in this case, eq. (63) is linear, while eq. (58) was not linear.

As for the finite difference approach, one may use the sensitivities u' and integrate them into the Gauss–newton-type methods, or simply use the cost function gradient, and then use the methods that do not rely on the sensitivities.

When compared to the finite difference approach, the forward difference method leads to exact cost function gradient components. Moreover, though $\mathcal{S}$ is likely to be a nonlinear operator, the system given by eq. (63) is linear, thus yielding to much less CPU time. Another singularity is that the discrete version of $\mathcal{S}'_u(u, \psi)$, i.e. $\mathcal{R}'_u$, is the tangent matrix that is to be used anyway for solving the "forward" problem

Algorithm 7: The finite difference algorithm to compute the gradient of the cost function and the sensitivities

Set the step $\varepsilon > 0$;
At iteration p, compute the state $u(\psi^p)$, compute $j(\psi^p)$;
foreach $i = 1, \dots, \dim \psi$ **do**
  Compute the perturbed state $u(\psi^p + \varepsilon|\psi_i|\delta\psi_i)$ and the cost $j(\psi^p + \varepsilon|\psi_i|\delta\psi_i)$;
  Set the state sensitivity $u'(\psi; \delta\psi_i) \leftarrow \dfrac{u(\psi^p + \varepsilon|\psi_i|\delta\psi_i) - u(\psi^p)}{\varepsilon|\psi_i|}$;
  Set the gradient $(\nabla j, \delta\psi_i)$ with either $(u - u_d, u'(\psi; \delta\psi_i))$ or as in previous algorithm with $\dfrac{j(\psi^p + \varepsilon|\psi_i|\delta\psi_i) - j(\psi^p)}{\varepsilon|\psi_i|}$.
end
Integrate the gradient within the optimization methods that do not rely on the sensitivities (conjugate gradient or BFGS among the presented methods) or within optimization methods that do rely on the sensitivities (Gauss–Newton or Levenberg–Marquardt).

$\mathcal{S}(u, \psi) = 0$. The computation of this linear tangent matrix is most often the task that takes the longer time in solving $\mathcal{S}(u, \psi) = 0$. The optimized procedure is thus the one given in algorithm 8.

Algorithm 8: The forward differentiation algorithm to compute the cost gradient and the sensitivities

At iteration p, solve iteratively $\mathcal{S}(u, \psi^p) = 0$, compute $j(\psi^p)$ and save the discrete version of the linear tangent operator $\mathcal{S}'_u(u, \psi^p)$;
foreach $i = 1, \dots, \dim \psi$ **do**
  Solve $\mathcal{S}'_u(u, \psi)u' + \mathcal{S}'_\psi(u, \psi^p)\delta\psi_i = 0$;
  Set $(\nabla j, \delta\psi_i)_{\mathcal{Z}} = (u - u_d, u'(\psi; \delta\psi_i)_{\mathcal{X}}$;
end
Integrate the gradient within the optimization methods that do not rely on the sensitivities (conjugate gradient or BFGS among the presented methods) or within optimization methods that do rely on the sensitivities (Gauss–Newton or Levenberg–Marquardt).
Note: the linear tangent matrix which is to be assembled for the solution of the "forward" model can be re-used for all canonical components $\delta\psi_i$.

Remark. *Equation (63) is often called the sensitivity equation.*

Example. *Let us consider the unsteady heat conduction equation, with known heat capacity C, conductivity λ, volume source term f, initial condition u_0, and Dirichlet condition on a part of the boundary, e.g. u_0 on ∂D_D. The unknown ψ is the flux ϕ on the rest of the boundary, i.e. on $\partial D_N = \partial D \backslash \partial D_D$.*
The unperturbed and perturbed models are:

$$\mathcal{S}(u,\psi) \equiv \begin{cases} C\frac{\partial u}{\partial t} - \nabla \cdot \lambda \nabla u = f & \text{in } D \\ u(x, t=0) = u_0 & \text{in } D \\ u(x,t) = u_0 & \text{on } \partial D_D \\ -\nabla u(x,t) \cdot n = \phi & \text{on } \partial D_N \end{cases} ; \mathcal{S}(u^+, \psi + \epsilon\delta\psi) \equiv \begin{cases} C\frac{\partial u^+}{\partial t} - \nabla \cdot \lambda \nabla u^+ = f & \text{in } D \\ u^+(x, t=0) = u_0 & \text{in } D \\ u^+(x,t) = u_0 & \text{on } \partial D_D \\ -\nabla u^+(x,t) \cdot n = \phi + \epsilon\delta\psi & \text{on } \partial D_N \end{cases} \tag{64}$$

Substracting the equations involved in these two models, dividing by ϵ, and searching the limit when $\epsilon \to 0$

gives:

$$\lim_{\epsilon\to 0}\frac{\mathcal{S}(u^+,\psi+\epsilon\delta\psi)-\mathcal{S}(u,\psi)}{\epsilon}\equiv\begin{cases}\lim_{\epsilon\to 0} C\frac{\partial\frac{u^+-u}{\epsilon}}{\partial t}-\nabla\cdot\lambda\nabla\frac{u^+-u}{\epsilon}=0 & \text{in } D\\ \lim_{\epsilon\to 0}\frac{u^+-u}{\epsilon}(x,t=0)=0 & \text{in } D\\ \lim_{\epsilon\to 0}\frac{u^+-u}{\epsilon}(x,t)=0 & \text{on } \partial D_D\\ \lim_{\epsilon\to 0}-\nabla\frac{u^+-u}{\epsilon}(x,t)\cdot n=\delta\psi & \text{on } \partial D_N\end{cases}\tag{65}$$

that gives:

$$\mathcal{S}'_u(u,\psi)u'+\mathcal{S}'_\psi\delta\psi\equiv\begin{cases}C\frac{\partial u'}{\partial t}-\nabla\cdot\lambda\nabla u'=0 & \text{in } D\\ u'(x,t=0)=0 & \text{in } D\\ u'(x,t)=0 & \text{on } \partial D_D\\ -\nabla u'(x,t)\cdot n=\delta\psi & \text{on } \partial D_N\end{cases}\tag{66}$$

6.3 Adjoint state

In this section we present the use of an additional problem – the so-called adjoint-state problem – that gives also the exact cost function gradient, but in a computational cheap way. We present one method based on the identification procedure (section 6.3.1), and another one that uses the Lagrange function (section 6.3.2). For the latter method, the model equation is treated as an equality constraint for the optimization. Both methods can deal with either the continuous equations or the discrete ones. One has to keep in mind that when the continuous method is used, all the obtained equations have later on to be discretized. Both strategies are equivalent in usual, but if the cost is computed through the integration of some discretized equations, then we consider that the discretized equations have to be differented (it is the so-called "discretize-then-differentiate" method). The other way is to deal with the continuous equations, then discretize the state model, etc. (it is the so-called "differentiate-then-discretize" method). Some examples of adjoint derivation will be given in the last sections.

6.3.1 Identification method

In this first part, we derive the adjoint-state method using the identification method. From the definition of the functional gradient, one writes the gradient:

$$(\nabla j,\delta\psi)_{\mathcal{Z}}=j'(\psi;\delta\psi)=\big(u-u_d,u'(\psi;\delta\psi)\big)_{\mathcal{X}}\,.\tag{67}$$

One then introduces a new variable (the adjoint-state variable u^*) such that the gradient equation given by eq. (67) also satisfies the "easier–to–compute":

$$j'(\psi;\delta\psi)=\big(\mathcal{S}'_\psi(u,\psi)\delta\psi,u^*\big)_{\mathcal{U}}\,.\tag{68}$$

On the other hand, since we have the relationship $\mathcal{S}(u,\psi)=0$, then

$$\mathcal{S}'_u(u,\psi)u'+\mathcal{S}'_\psi(u,\psi)\delta\psi=0\tag{69}$$

and thus, we have:

$$j'(\psi;\delta\psi)=-\big(S'_u(u,\psi)u',u^*\big)_{\mathcal{U}}\,.\tag{70}$$

Identifying eq. (67) and eq. (70), we obtain the adjoint-state problem that must satisfy the equality:

$$-\big(\mathcal{S}'_u(u,\psi)u',u^*\big)_{\mathcal{U}}=\big(u-u_d,u'(\psi;\delta\psi)\big)_{\mathcal{X}}\,.\tag{71}$$

Next, if the adjoint problem given by eq. (71) is satisfied (it means that we accessed the adjoint state u^*), then the cost function gradient is very simply given by eq. (68). We then use the inner product property

$(\mathscr{A}u, v) = (u, \mathscr{A}^* v)$ where $\mathscr{A}^*$ is the transposed conjugate operator of $\mathscr{A}$ (adjoint) to modify the adjoint equation given by eq. (71) to:

$$\mathcal{S}^*(u, \psi)u^* + (u - u_d) = 0, \tag{72}$$

where $\mathcal{S}^*$ is the conjugate transposed of the linear tangent operator $\mathcal{S}'_u$, i.e., we used:

$$\left(\mathcal{S}'_u(u, \psi)u', u^*\right)_{\mathcal{U}} = \left(S^*(u, \psi)u^*, u'\right)_{\mathcal{U}} + [\cdots] \tag{73}$$

where the term $[\cdots]$ may contain some additional terms coming from some integrations by parts. Figure 11 schematically represents the process of identification method.

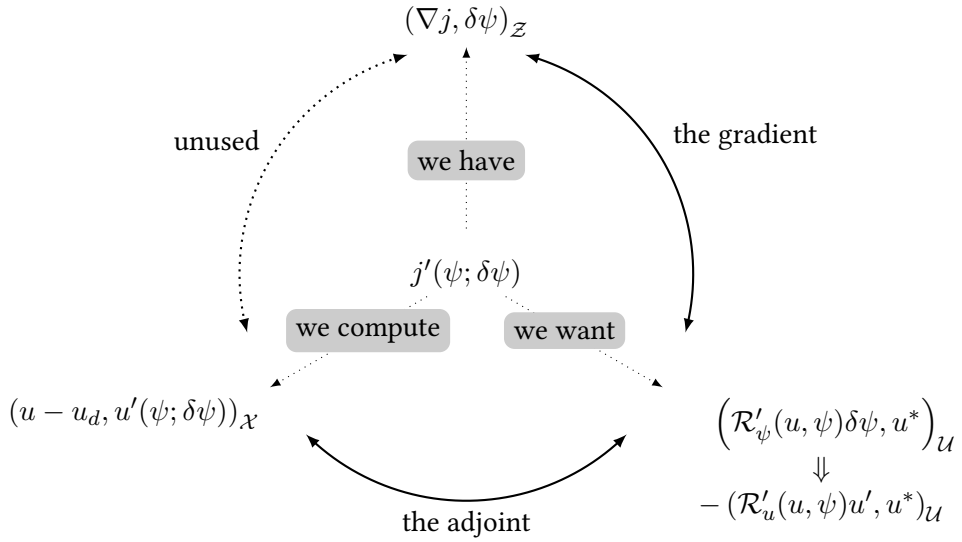


Figure 11: Schematical representation of the adjoint-state method.

Remark. *The inner product $(v, w)_{\mathcal{U}}$ is performed on the whole domain of definition of u. For instance if $u \in L_2(0, T; L_2(\mathcal{D}))$, then $(v, w)_{\mathcal{U}} = \int_0^T \int_{\mathcal{D}} vw \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t$.*

Algorithm 9: The adjoint state problem to compute the cost function gradient with integration within an optimization algorithm

At iteration p, solve iteratively $\mathcal{S}(u, \psi) = 0$;
Compute $j(\psi^p)$;
Save the solution u;
Compute the adjoint state problem $\mathcal{S}^*(u, \psi)u^* + (u - u_d) = 0$;
Compute the gradient $(\nabla j(\psi); \delta\psi)_{\mathcal{Z}} = \left(\mathcal{S}'_\psi(u, \psi)\delta\psi, u^*\right)_{\mathcal{U}}$;
Integrate the gradient within the optimization methods that do not rely on the sensitivities (conjugate gradient or BFGS among the presented methods)

6.3.2 Lagrange formulation

The use of a Lagrange formulation means that the state equations are taken as constraints in the optimization problem. To do so, let us introduce the Lagrange function [12, 13]:

$$\mathscr{L}(u, u^*, \psi) = \mathcal{J}(u) + (\mathcal{S}(u, \psi), u^*)_{\mathcal{U}} \tag{74}$$

The Lagrange function introduced in this section is a function of three variables, namely the state u , the parameter to be to be identified ψ, and the adjoint state variable u^*. This means that both variables u and ψ are somehow considered to be independent, even though there exists, at least implicitly, the relationship $\mathcal{S}(u,\psi)=0$ that maps ψ to u. Moreover, since u is the solution of the forward model, then the Lagrange function $\mathscr{L}$ is always equal to the cost function $\mathcal{J}(u)$, and the constraints – which represent the partial differential equations of the forward problem – are always satisfied. We now show that a necessary condition for the set ψ to be solution of the optimization problem eq. (1) is that there exists a set (u,ψ) such that (u,ψ,u^*) is a saddle point (stationary point) of $\mathscr{L}$. Indeed, let us show that the necessary condition $j'(\psi;\delta\psi)=0, \forall\delta\psi$, is equivalent to:

$$\exists\,(u,u^*,\psi)\ \mid \mathscr{L}'_u\,(\cdot)\,\delta w=0;\ \mathscr{L}'_{u^*}\,(\cdot)\,\delta w=0;\ \mathscr{L}'_{\psi}\,(\cdot)\,\delta w=0, \tag{75}$$

for all directions δw taken in appropriate spaces (u', δu^* and $\delta\psi$). First, since the state is satisfied, then:

$$\mathscr{L}'_{u^*}=\mathcal{S}(u,\psi)=0.$$

Moreover, since we have also $\mathcal{S}'_u(u,\psi)u'+\mathcal{S}'_{\psi}(u,\psi)\delta\psi=0$, we get:

$$\mathscr{L}'_{\psi}\,(\cdot)\,\delta\psi=\left(\mathcal{S}'_{\psi}(u,\psi)\delta\psi,u^*\right)_{\mathcal{U}}=-\left(\mathcal{S}'_u(u,\psi)u',u^*\right)_{\mathcal{U}}. \tag{76}$$

In another hand, the differentiation of the Lagrange function with respect to the state gives:

$$\mathscr{L}'_u\,(\cdot)\,u'=\left(u-u_d,u'\right)_{\mathcal{X}}+\left(\mathcal{S}'_u(u,\psi)u',u^*\right)_{\mathcal{U}}. \tag{77}$$

So far, the choice for the adjoint variables u^* has not been fixed yet. However, choosing the adjoint variable such that $\mathscr{L}'_u\,(\cdot)\,u'=0\ \forall u'$ considerably simplifies the relationship between the differentiated lagrangian with respect to ψ and the cost function gradient. One actually chooses u^* such that it satisfies the adjoint-state equation:

$$\left(S'_u(u,\psi)u',u^*\right)_{\mathcal{U}}+\left(u-u_d,u'(\psi;\delta\psi)\right)_{\mathcal{X}}=0. \tag{78}$$

This way we obtain the cost function gradient:

$$\mathscr{L}'_{\psi}\,(\cdot)\,\delta\psi=\left(u-u_d,u'(\psi;\delta\psi)\right)_{\mathcal{X}}=j'(\psi;\delta\psi)=(\nabla j,\delta\psi)_{\mathcal{Y}} \tag{79}$$

The adjoint-state equation is thus:

$$\mathcal{S}^*(u,\psi)u^*+(u-u_d)=0, \tag{80}$$

and the gradient is given by:

$$\nabla j=\left(\mathcal{S}'_{\psi}(u,\psi),u^*\right)_{\mathcal{Y}}. \tag{81}$$

Summarizing, the minimum of the cost function is to be found at the stationary point of the Lagrange function eq. (74). When the adjoint-state equation eq. (80) is satisfied, then the components of the cost function gradient are simply given through the inner product eq. (81).

6.3.3 Examples

In the examples presented below, we do not specify what the parameters are. We just give the form of the adjoint-state problem related to the "forward" state problem form.

Case of ODE Let us start with the case where the state model is simplified to a single linear continuous ordinary differential equations integrated in time $\mathcal{I} = (0, t_f]$. The forward problem thus writes:

$$\begin{aligned} \mathcal{S}(u, \psi) = \mathscr{C}\dot{u} - \mathscr{B} = 0 \qquad & \text{for } t \in \mathcal{I} \\ u = u_0 \qquad & \text{for } t = 0, \end{aligned} \tag{82}$$

where $\mathscr{C}$ is an inertial scalar term and $\mathscr{B}$ contains the loadings. Injecting the differentiated time-dependent relationship eq. (82) into the adjoint-state relationship eq. (78) gives:

$$\left(\mathscr{C}\dot{u}', u^*\right)_{\mathcal{U}} + \left(u - u_d, u'\right)_{\mathcal{X}} = 0$$

where the inner must be understood as $(a, b)_{\mathcal{U}} = \int_{\mathcal{I}} ab \, \mathrm{d}t$. One then integrates by part the first term to get:

$$-\left(u', \mathscr{C}\dot{u}^*\right)_{\mathcal{U}} + \left[u'\mathscr{C}u^*\right]_0^{t_f} + \left(u - u_d, u'\right)_{\mathcal{X}} = 0$$

Since there is no reason that the initial state depend on the parameters ψ (except if the initial state is searched), then the derivatives u' of u at initial time is zero. The adjoint-state problem is eventually:

$$\begin{aligned} -\mathscr{C}\dot{u}^* + (u - u_d) = 0 \qquad & \text{for } t \in \mathcal{I} \\ \mathscr{C}u^* + (u - u_d) = 0 \qquad & \text{for } t = t_f \ . \end{aligned} \tag{83}$$

Remark. *There is a minus sign just before the operator $\mathscr{C}$ involved in the first equation. At the same time, the boundary-time condition is given at final time t_f. Therefore, when considering these two points, there is no way to solve the adjoint problem forwardly, i.e., from $t = 0$ to t_f. The trick consists in introducing a new time variable $\tau = t_f - t$ (the dual time). Doing so, the initial condition is given at the initial time $\tau = 0$, and the time-dependent equation eq. (83) is solved in the forward way in the dual time variable τ – which corresponds to the backward way in the primal time variable t.*

Remark. *The loading component $(u - u_d)$ involved in eq. (83) is non-zero only at times where the cost function j is to be integrated, i.e., in accordance with the definition of the $\mathcal{X}$-norm.*

Remark. *Inherently, the adjoint-state problem is linear: even though the forward problem is likely to be nonlinear (it was not the case in the considered exemple), the adjoint-state problem is still linear since the operators do not depend on the adjoint-state variables. An equivalent remark was given for the forward differentiation method which used the linear tangent operator.*

Case of elliptic PDE This second example concerns the case where the state model is simplified to a diffusive-type continuous partial differential equation independent of time:

$$\mathcal{S}(u, \psi) = -\nabla \cdot \lambda \nabla u - f = 0 \quad \text{in } \mathcal{D}. \tag{84}$$

Injecting the differentiated space-dependent relation eq. (84) into the adjoint eq. (78) gives:

$$\left(-\lambda \Delta u', u^*\right)_{\mathcal{U}} + \left(u - u_d, u'\right)_{\mathcal{X}} = 0.$$

with $(a, b)_{\mathcal{U}} = \int_{\mathcal{D}} ab \, \mathrm{d}x$. Using twice the Green theorem on the first integral, one gets:

$$\left(u', -\lambda \Delta u^*\right)_{\mathcal{U}} + (\ldots)_{\partial\mathcal{U}} + \left(u - u_d, u'\right)_{\mathcal{X}} = 0. \tag{85}$$

Owing to be verified for all directional derivatives u', the general adjoint-state problem becomes:

$$-\lambda \Delta \nabla u^* + (u - u_d) = 0. \tag{86}$$

Remark. *The second term, $(\ldots)_{\partial\mathcal{U}}$ comes in eq. (85) because of the integration by parts. These terms depend on the boundary conditions associated to eq. (84) that formed the complete forward model. Taking into account of these terms will also complete the definition of the adjoint-state model, yielding the boundary conditions associated to the adjoint-state equation eq. (86).*

Remark. *The loading component $u - u_d$ involved in the space-dependent equation is non-zero only at the selected locations where the cost function j is to be integrated, i.e., in accordance with the definition of the $\mathcal{X}$-norm.*

Case of parabolic PDE The discretization of the space and time dependent diffusive model yields to the so-called parabolic problem. It is somehow the union between both just above presented cases:

$$\begin{array}{ll} \mathcal{S}(u,\psi) = \mathscr{C}\dot{u} - \Delta u - \mathscr{B} = 0 & \text{for } t \in \mathcal{I},\ x \in \mathcal{D} \\ u = u_0 & \text{for } t = 0, \end{array} \tag{87}$$

with associated boundary conditions. Injecting the differentiated operators involved in eq. (87) into the adjoint eq. (78) gives:

$$\left(\mathscr{C}\dot{u}', \psi\right)_{\mathcal{U}} - \left(\Delta u', \psi\right)_{\mathcal{U}} + \left(u - u_d, u'\right)_{\mathcal{X}} = 0$$

with $(a,b)_{\mathcal{U}} = \int_{\mathcal{T}} \int_{\mathcal{D}} ab \,\mathrm{d}x \,\mathrm{d}t$. Transposing all operators through integration by parts (once in time and twice space) gives:

$$-\left(u', \mathscr{C}\dot{u}^*\right)_{\mathcal{U}} + \left[\left(u', \mathscr{C}u^*\right)_{\mathcal{D}}\right]_0^T - \int_{\mathcal{I}} [...]_{\partial\mathcal{D}} \,\mathrm{d}t + \left(u - u_d, u'(\psi; \delta\psi)\right)_{\mathcal{X}} = 0$$

Eventually, the adjoint problem becomes:

$$\begin{array}{ll} -\mathscr{C}\dot{u}^* - \Delta u^* + \mathcal{J}'(u) = 0 & \text{for } t \in \mathcal{I} \\ u^* = 0 & \text{for } t = t_f. \end{array} \tag{88}$$

along with associated spatial boundary conditions.

Remark. *A more detailed example given later on in section 9.2 provides the full calculation of the boundary condition for a similar case.*

6.4 The global optimization algorithm

The general algorithm is given in algorithm 10. The global procedure described in this algorithm is run until (at least) one of the stopping criteria presented in section 2.5 is reached.

Algorithm 10: The global optimization algorithm

1. Integrate the cost function value through integration of the forward (maybe nonlinear) problem;
   Store all state variables to reconstruct the tangent matrix (or store the tangent matrix);
2. Integrate the backward linear adjoint-state problem, all matrices being possibly stored or recomputed from stored state variables
3. Compute the cost function gradient;
   Compute the direction of descent
4. Solve the line-search algorithm

6.5 Continuous gradient and discretized continuous gradient

In previous examples as well as in the derivation of both forward differentiated and adjoint-state models, all derivations were performed on continuous equations. To be solved, such equations will have later on to be discretized, for example with finite elements or any other method. This ordinary process yields the so-called *continuous gradient*. For example, refering to fig. 12, the continuous state equation for the continuous variable u is first differentiated, yielding a continuous differentiated state u', solution of a continuous differentiated partial differential equation. Then, after discretization (which is an approximation process), one has the discretized differentiated state $(u')_h$, such that the discretized continuous gradient $j'|_{\mathrm{DCG}}$ can be computed.

The other way round consists in first discretizing, then differentiating. All partial differential equations are discretized, so that the forward state is u_h, and the corresponding cost function, based on this approximated state is j_h. This approximated model can then be differentiated so that the derivative we get is $(u_h)'$, which gives the discrete gradient $j'|_{\text{DG}}$.

Both gradients $j'|_{\text{DCG}}$ and $j'|_{\text{DG}}$ are different because the approximations are not performed on the same operators. The cost function is always j_h because a numerical solver is used to compute u_h. The minimum of j_h corresponds to $j'|_{\text{DG}} = 0$ but at this minimum, it is likely that we have $j'|_{\text{DCG}} \neq 0$. This means that the discrete gradient is compatible with the cost function which is calculated while the discretized continuous gradient is not. However, one has to keep in mind that errors coming from discretization are likely to be negligible when compared to measurement errors of the inverse problem. As such, the computation of discretized continuous gradients is – according to the author – the better strategy because all derivations are performed on partial differential equations, and differentiated models are also partial differential equation very similar to model equation, and such similarity is the easiest way to go: similar equation, re-use the forward solver for the differentiated model or for the adjoint-state, etc.

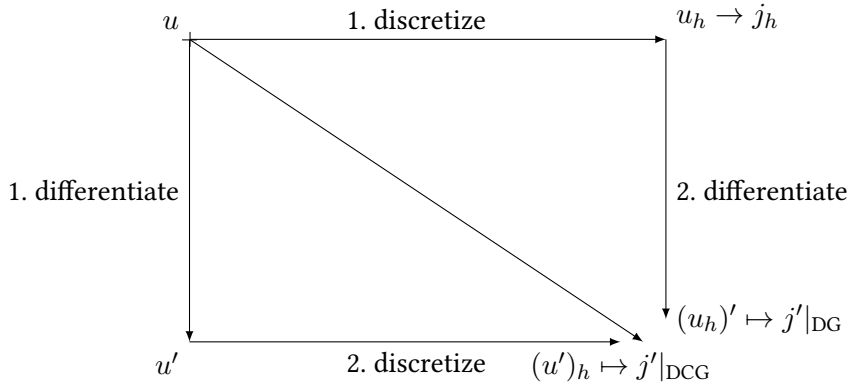


Figure 12: Discrete gradient vs discretized continuous gradient

7 Elements of comparison

We give in this section some elements of comparison between the previously presented optimization algorithms and between the different gradient computation strategies.

7.1 Convergence speed

The optimization algorithms presented in section 5 yield to a series $\{\psi^k\}_{k\geq 1}$ that converges to $\bar{\psi}$. Hereafter are some some convergence rate definitions [3, 2].

Definition 7. *The convergence rate of the series $\{\psi^k\}_{k\geq 1}$ is said to be linear if*

$$\frac{\|\psi_{k+1} - \psi_k\|}{\|\psi_k - \bar{\psi}\|} \leq \tau, \quad \tau \in (0, 1). \tag{89}$$

This means that the distance to the solution $\bar{\psi}$ decreases at each iteration by at least the constant factor τ.

Definition 8. *The convergence rate of the series $\{\psi^k\}_{k\geq 1}$ is said to be superlinear in n steps if*

$$\lim_{k\to\infty} \frac{\|\psi_{k+n} - \psi_k\|}{\|\psi_k - \bar{\psi}\|} = 0. \tag{90}$$

Definition 9. *The convergence rate of the series* $\{\psi^k\}_{k\geq 1}$ *is said to be quadratic if*

$$\frac{\|\psi_{k+1}-\psi_k\|}{\|\psi_k-\bar{\psi}\|^2}\leq \tau,\quad \tau>0. \tag{91}$$

Quasi–Newton methods usually converge super-linearly and the Newton method converges quadratically. The steepest descent method converge linearly. Moreover, for ill-posed problems, this method may converge linearly with a constant τ close to 1. Next, the conjugate-gradient method converges superlinearly in n steps to the optimum [2].

Thus the quasi-Newton methods convergence-rate is much higher than the conjugate gradient methods convergence-rate which need approximately n times more steps (n times more line-search) at the same convergence behavior. However, for the quasi-Newton method, the memory place is proportionnal to n^2.

7.2 Gradient computation cost

Let $\mathcal{S}(u,\psi)=0$ the state problem that maps $\psi\mapsto u$, $\mathcal{R}$ being possibly nonlinear (for highlighting differences between the distinct strategies), and $\dim\psi$ the number of parameters to be evaluated. We compare the number of times the model $\mathcal{S}$, the differentiated model and/or the adjoint-state model are computed to access the full gradient of the cost function.

1. Finite difference method:
   $(\dim\psi+1)$ nonlinear computation of $\mathcal{S}(u,\psi)=0$.

2. Forward differentiation method:
   1 nonlinear computation of $\mathcal{S}(u,\psi)=0$,
   $\dim\psi$ linear computation of $\mathcal{S}'_u(u,\psi)u'+\mathcal{S}'_\psi(u,\psi)\delta\psi=0$.

3. Adjoint state method:
   1 nonlinear computation of $\mathcal{S}(u,\psi)=0$,
   1 linear computation of $\mathcal{S}^*(u,\psi)u^*+u-u_d=0$.

Thus, the finite difference method is very time consuming, though it is easy to use. Next, comparing the two latter methods, the operator involved in the adjoint-state method is almost the same as the one involved in the forward differentiation method, though the adjoint-state method yields to higher algorithmic complexity (backward time integration, memory, etc.). When $\dim\psi$ is high (even if $\dim\psi$ is bigger than say 100), the use of the direct differentiation method becomes cumbersome and computationally expensive; the adjoint-state method is, in fact, the only acceptable method.

7.3 Gradient computation needs

We recall in the following table the way (the required needed steps) one computes the cost function gradient.

Steepest, conjugate-gradients, BFGS, DFP, …	Newton	Gauss–Newton, Levenberg–Marquardt, …
$u\leftarrow\mathcal{S}(u,\psi)=0$ $j\leftarrow u$ $\nabla j\leftarrow$ { Forward diff. or Adjoint state }	$u\leftarrow\mathcal{S}(u,\psi)=0$ $j\leftarrow u$ $\nabla j\leftarrow$ { Forward diff. or Adjoint state } $\nabla^2 j$ (complicated)	$u\leftarrow\mathcal{S}(u,\psi)=0$ $j\leftarrow u$ $\nabla j\leftarrow S^tS\leftarrow S\leftarrow u'$ (Forward diff.)

8 Regularization

When the inverse problem is ill-posed (which is likely to be the case in real cases, especially when the control space dimension is big), regularization is needed and sometimes compulsary. Regarding function estimation, for instance space-dependent physical properties or sources, specific regularization strategies different from the ones used in parametric estimation are required. Regularization may be viewed as adding *a priori* information, but other means can also be used, including (see [14] for elements of comparison on applications of optical tomography):

- choose of specific $\mathcal{X}$-norm for the cost function expression according to the prior knowledge of the unknown (use for instance the $L_1(\mathcal{D})$-norm instead of the ordinary $L_2(\mathcal{D})$-norm).
- add prior information through Tikhonov penalization, If some Tikhonov-type regularization terms are added to the cost function, the cost function $j_\epsilon(\psi) := \mathcal{J}(u) + \epsilon\mathcal{J}^+(\psi)$ is the one to be minimized, with:
$$\mathcal{J}^+ := \|\mathbf{D}\psi\|^2_{\mathcal{Y}} \tag{92}$$
where $\mathbf{D}$ is often a differential operator acting on the function ψ and $\|\cdot\|_{\mathcal{Y}}$ is another norm to be defined according to the chosen control space.
- choose an appropriate $\mathcal{Z}$-norm for extracting the cost function gradient. The use of specific inner products when extracting the cost function gradient is a recent regularization tool. In order to present this regularization strategy, let us work on the example where a space-dependent physical property in $\mathcal{D}$ is to be estimated. In such a case it is usual to use the ordinary $L_2(\mathcal{D})$-inner product, i.e., one uses $j'(\psi;\phi) = (\nabla j, \phi)_{L_2(\mathcal{D})}$; this gives the ordinary $L_2(\mathcal{D})$ cost function gradient, denoted here as $\nabla^{L_2} j(\psi)$. Besides, the Sobolev inner product can give much better (smoother) results when the noise has propagated to the adjoint-state variable and then to the cost function gradient. Even better, the weighted version has recently proven to give excellent results. This one defined as:
$$(u,v)_{\mathcal{Z}} = (u,v)_{H^{1(\ell)}(\mathcal{D})} := \int_{\mathcal{D}} \left(uv + \ell^2 \nabla u \cdot \nabla v\right) \, \mathrm{d}\boldsymbol{x} \tag{93}$$
is used in the cost function gradient extraction relationship $j'(\psi;\phi) = (\nabla j, \phi)_{H^{1(\ell)}(\mathcal{D})}$ in order to compute the weighted Sobolev cost function gradient $\nabla^{H^{1(\ell)}} j(\psi)$.
- choose an appropriate functional space for the control space parameterization. In practice, the control space must be approximated in order to be finite. Often, the finite element method is used so that one searches ψ that belongs to a finite dimensional subspace, say $\mathcal{V}$. Let us consider a triangulation $\mathcal{M}$ of the computational domain $\mathcal{D}$, and let us note n_p the number of vertices in $\mathcal{M}$. It has been shown, through numerical means on a specific OT problem that, among the large number of tested possibilities, the piecewise linear continuous functions ($\dim \psi = n_p$) are the most appropriate for the estimation of space-dependent functions.
- choose an appropriate dimension $\dim \psi$ of the control space parameterization. Usually the finite element space used to solve the forward model (58) has to be fine enough to ensure that numerical errors stay small enough. Most often, the triangulation chosen for the control space is the one chosen for the state. It has been shown again, through numerical means, that both the convergence and the quality of the reconstructions are much improved when $\dim \psi$ is lowered, up to a certain limit, at least for quasi-Newton algorithms.
- Multi-scales approaches is also a fabulous opportunity to regularize solutions and in the same time accelerate the convergence and avoid converging to local minima. Coupled with wavelets on one

side, and the BFGS in the other side, this method relies on a reformulation of the original inverse problem into a sequence of sub-inverse problems of different scales using wavelet transform, from the largest scale to the smallest one. Successful applications of this method include the estimation of space-dependent absorption and scattering coefficients in optical tomography [15].

9 Examples

9.1 Parametric conductivities in a transient heat conduction problem

This first simple example deals with the estimation of uniform conductivity coefficients in different subdomains. Heat transfer is considered. Initial temperature is assumed to be known and equal to T_0. T_0 is also the Dirichlet temperature for positive time on the whole boundary $\partial\mathcal{D}$. The domain has the shape of a head with two eyes, one nose and one mouth. The geometry being known, as well as the initial and boundary conditions, the heat capacity and the time-dependent heat source, the inverse heat conduction problem consists in estimating, through infra-red like temperature measurements on $\mathcal{D}\times\mathcal{I}$, the set of conductivities λ_i, $i=1,\cdots,4$ (1, 2, 3, 4 corresponding to the left eye, the right eye, the noze and the mouth, respectively). Noisy (1 % white noise) synthetic data was generated with conductivities equal to 20, 30, 40 and 50, respectively. Guessed conductivities were all equal to 10.

If optimization methods based on sensitivities are chosen, one will have to compute, sucessively:

$$\rho C\frac{\partial T'}{\partial t}-\nabla\cdot\left(\lambda T'\right)=\nabla\cdot\left(\lambda'\nabla T\right) \tag{94}$$

with null initial and boundary conditions. In the sensitivity model, the direction λ' equals 0 or 1 depending on the location for the four considered sensitivity problems. Corresponding sensitivities are presented in fig. 13.

With so few parameters to identify (4 in total in this example), it is not really necessary to use the adjoint-state method to compute the cost gradient. We however give in fig. 13 the evolution of the adjoint-state variable which is solved backwardly from final time to initial time, while integrating along time the errors integrated within the cost function (this first application was actually chosen for this purposes : small number of unknowns, and possible visualization).

From the knowledge of these temperature sensitivities, one can compute the sensitivity matrix S such that $s_{i\times k,j}$ gathers for instance the sensitivity of temperature on the (finite element) node i at time k with respect to λ_j. In the same manner, the error vector $e_{i\times k}$ gathers the error (difference between the predictions and the measurements) at the (FE) node i, and at time t_k. Consequently, the cost gradient is computed straightforwardly through $\nabla j=S^\top e$, and the Gauss–Newton algorithm can be used without any regularization because this parametric problem is not ill-posed. Very few Gauss–Newton iterations are needed to converge as can be seen in fig. 14.

9.2 Space-dependent convection coefficient in a transient heat conduction problem

In this section, we consider an application of a nonlinear transient heat transfer inverse problem arising in thermal treatment for instance. $\mathcal{D}$ being an open bounded set of $\mathbb{R}^2$ and $\mathcal{I}=]0,t_f]$, the modeling equation in $\mathcal{D}\times\mathcal{I}$ is

$$C\frac{\partial T}{\partial t}-\nabla\cdot(\lambda\nabla T)=f\quad\text{ for }(x,t)\in\mathcal{D}\times\mathcal{I} \tag{95}$$

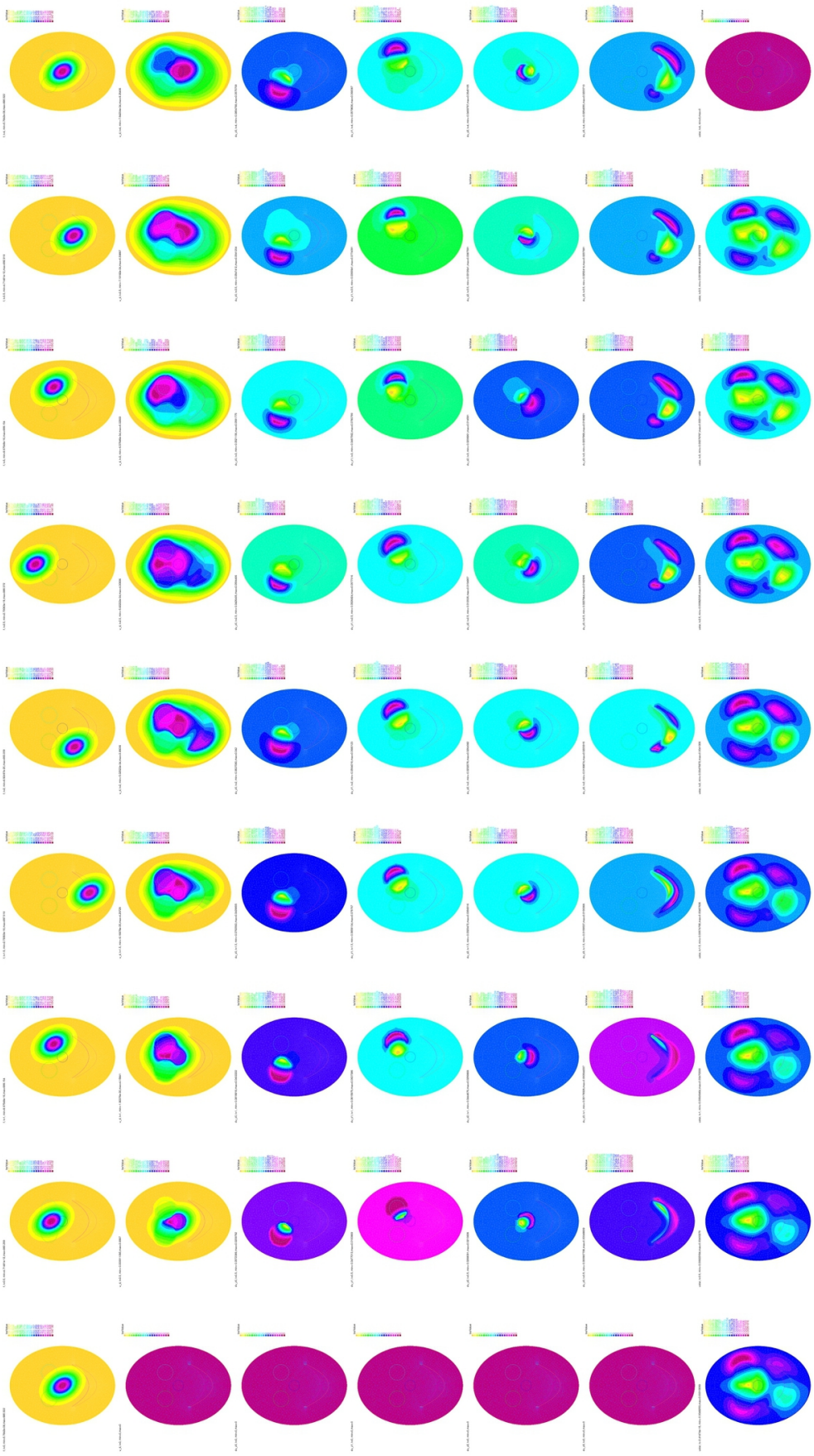


Figure 13: Variables needed to solve the parametric inverse transient heat conduction problem. Columns correspond to increasing time from 0 to 40 by steps of 5. The first raw presents the source term that follows, in this particular case, a lemniscate. The second raw presents the temperature evolution. The four following raws present the evolution of sensitivities with respect to λ_1 (the left eye), λ_2 (the right eye), λ_3 (the noze), and λ_4 (the mouth), respectively. The last raw presents the evolution of the adjoint-state: it is null at final time and then increases due to cost function integration while going back to initial time.

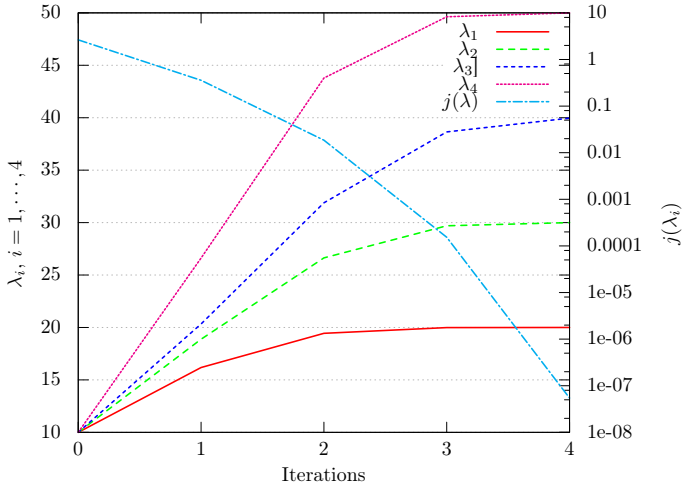


Figure 14: Evolution of the parameters and of the cost function with respect to the G-N iterations.

where the temperature-dependent physical properties are considered. We also consider the following initial and boundary conditions with $\partial\mathcal{D}_1 \oplus \partial\mathcal{D}_3 \oplus \partial\mathcal{D}_3$ forming a partition of $\partial\mathcal{D}$:

$$\begin{array}{rll} T & = T_0 & \text{at } t = 0 \\ -\lambda\nabla T \cdot \boldsymbol{n} & = h\left(T - T_\infty\right) & \text{for } \boldsymbol{x} \in \partial\mathcal{D}_1 \\ \nabla T \cdot \boldsymbol{n} & = 0 & \text{for } \boldsymbol{x} \in \partial\mathcal{D}_2 \\ -\lambda\nabla T \cdot \boldsymbol{n} & = \varepsilon\sigma\left(T^4 - T_\infty^4\right) & \text{for } \boldsymbol{x} \in \partial\mathcal{D}_3 \end{array} \tag{96}$$

The estimation of the heat transfer function $h(\boldsymbol{x}, t)$, $\boldsymbol{x} \in \partial\mathcal{D}_1$, $t \in]0, t_f]$ is performed through the minimization of the cost function:

$$j(h) := \mathcal{J}(T) = \int_0^{t_f} \sum_{j=1}^{N} (T(\boldsymbol{x}_j, t) - T_d(\boldsymbol{x}_j, t))^2 dt \tag{97}$$

where $T(\boldsymbol{x}_j, t)$ and $T_d(\boldsymbol{x}_j, t)$ represent respectively the predicted and measured temperatures at N various locations $\boldsymbol{x} := (r_j, z_j)$ in the material. For such application, the minimization can be carried out by using conjugate gradients or better quasi-Newton methods. In any case, the optimization is based on the gradient computation.

The cost function gradient is obtained for all values $\boldsymbol{x} \in \partial\mathcal{D}_1$, $t \in]0, t_f]$ by the following relationship:

$$\nabla j(h) = T^* \times (T - T_\infty) \tag{98}$$

where T^* is the solution of the adjoint problem:

$$-C\frac{\partial T^*}{\partial t} - \nabla \cdot (\lambda \nabla T^*) = \sum_j (T - T_d) \times \delta\left(\boldsymbol{x} - \boldsymbol{x}_j\right) \tag{99}$$

with the condition $T^* = 0$ at final time t_f and the conditions on the boundaries:

$$\begin{array}{rll} -\lambda\nabla T^* \cdot \boldsymbol{n} = hT^* & & \text{for } \boldsymbol{x} \in \partial\mathcal{D}_1 \\ \nabla T^* \cdot \boldsymbol{n} = 0 & & \text{for } \boldsymbol{x} \in \partial\mathcal{D}_2 \\ -\lambda\nabla T^* \cdot \boldsymbol{n} = 4\varepsilon\sigma T^3 T^* & & \text{for } \boldsymbol{x} \in \partial\mathcal{D}_3 \end{array} \tag{100}$$

Remark. *The following notations are used:* $\mathcal{U} := L_2(\mathcal{I}; L_2(\mathcal{D}))$, $\mathcal{U}_i := L_2(\mathcal{I}; L_2(\partial\mathcal{D}_i))$, $\forall i = 1, 2, 3$ *and* $\mathcal{U}_* := L_2(\mathcal{I}; L_2(\cup_i \partial\mathcal{D}_i))$.

Proof. The derivative of the state T at the point h and towards δh, $T'(h;\delta h)$ is defined by:

$$\left\{\begin{array}{ll} C\frac{\partial T'}{\partial t} - \nabla\cdot\lambda\nabla T' = 0 & \boldsymbol{x}\in\mathcal{D},\, t\in\mathcal{I} \\ T' = 0 & \boldsymbol{x}\in\mathcal{D},\, t=0 \\ \lambda\nabla T'\cdot\boldsymbol{n} + hT' + \delta h(T-T_\infty) = 0 & \boldsymbol{x}\in\partial\mathcal{D}_1,\, t\in\mathcal{I} \\ \nabla T'\cdot\boldsymbol{n} = 0 & \boldsymbol{x}\in\partial\mathcal{D}_2,\, t\in\mathcal{I} \\ \lambda\nabla T'\cdot\boldsymbol{n} + 4\varepsilon\sigma T^3 T' = 0 & \boldsymbol{x}\in\partial\mathcal{D}_3,\, t\in\mathcal{I} \end{array}\right. \tag{101}$$

The Lagrange function is formally defined as:

$$\begin{array}{rl} \mathscr{L}(T,\{T^*,\gamma,\xi,\varpi\},h) = & \mathcal{J}(T) + \left(C\frac{\partial T}{\partial t} - \nabla\cdot(\lambda\nabla T) - f, T^*\right)_{\mathcal{U}} \\ & + \left(\lambda\nabla T\cdot\boldsymbol{n} + h\left(T-T_\infty\right),\gamma\right)_{\mathcal{U}_1} \\ & + \left(\nabla T\cdot\boldsymbol{n},\xi\right)_{\mathcal{U}_2} \\ & + \left(\lambda\nabla T\cdot\boldsymbol{n} + \varepsilon\sigma\left(T^4 - T_\infty^4\right),\varpi\right)_{\mathcal{U}_3} \end{array} \tag{102}$$

The differentiated Lagrange function with respect to h in the direction δh is:

$$\begin{array}{rl} (\mathscr{L}'_h(\cdot),\delta h) = & (T-T_d, T')_{\mathcal{X}} \\ & + \left(C\frac{\partial(T')}{\partial t} - \nabla\cdot\lambda\nabla T', T^*\right)_{\mathcal{U}} \\ & + (\lambda\nabla T')\cdot\boldsymbol{n} + hT' + \delta h(T-T_\infty),\gamma)_{\mathcal{U}_1} \\ & + (\nabla T'\cdot\boldsymbol{n},\xi)_{\mathcal{U}_2} \\ & + \left(\lambda\nabla T'\cdot\boldsymbol{n} + 4\varepsilon\sigma T^3 T', \varpi\right)_{\mathcal{U}_3} \end{array} \tag{103}$$

We then use the following integrations by parts to express some particular terms:

$$\begin{array}{rl} \left(C\frac{\partial T'}{\partial t}, T^*\right)_{\mathcal{U}} = & \left(T', -C\frac{\partial T^*}{\partial t}\right)_{\mathcal{U}} + (CT', T^*)_{\mathcal{D}}\left(t=t_f\right) - (CT', T^*)_{\mathcal{D}}\left(t=0\right) \\ (\lambda\Delta T', T^*)_{\mathcal{U}} = & (\lambda\Delta T^*, T')_{\mathcal{U}} + (\lambda\nabla T^*\cdot\boldsymbol{n}, T')_{\mathcal{U}_*} - (\lambda T^*, \nabla T'\cdot\boldsymbol{n})_{\mathcal{U}_*} \end{array} \tag{104}$$

We bring together similar terms to get:

$$\begin{array}{rl} (\mathscr{L}'_h(T,\{T^*,\gamma,\xi,\varpi\},h),\delta h) = & (T-T_d,T')_{\mathcal{X}} + \left(\delta h\left(T-T_\infty\right),\gamma\right)_{\mathcal{U}_1} \\ & + \left(-C\frac{\partial T^*}{\partial t} - \lambda\Delta T^*, T'\right)_{\mathcal{U}} + (CT^*, T')_{\mathcal{D}}\left(t=t_f\right) \\ & + (\lambda\nabla T^*\cdot\boldsymbol{n}, T')_{\mathcal{U}_*} - (\lambda T^*, \nabla T'\cdot\boldsymbol{n})_{\mathcal{U}_*} \\ & + (\lambda\nabla T')\cdot\boldsymbol{n} + hT', \gamma)_{\mathcal{U}_1} + (\nabla T'\cdot\boldsymbol{n},\xi)_{\mathcal{U}_2} \\ & + \left(\lambda\nabla T'\cdot\boldsymbol{n} + 4\varepsilon\sigma T^3 T', \varpi\right)_{\mathcal{U}_3} \end{array} \tag{105}$$

Choosing $\gamma = T^*$ on $\partial\mathcal{D}_1$, $\xi = \lambda T^*$ on $\partial\mathcal{D}_2$ and $\varpi = T^*$ on $\partial\mathcal{D}_3$, the adjoint problem can eventually be written as:

$$\left\{\begin{array}{ll} -C\frac{\partial T^*}{\partial t} - \lambda\Delta T^* = -\sum_j (T-T_d)\times\delta(\boldsymbol{x}-\boldsymbol{x}_j) & \boldsymbol{x}\in\mathcal{D},\, t\in\mathcal{I} \\ T^* = 0 & \boldsymbol{x}\in\mathcal{D},\, t=t_f \\ -\lambda\nabla T^*\cdot\boldsymbol{n} = hT^* & \boldsymbol{x}\in\partial\mathcal{D}_1,\, t\in\mathcal{I} \\ \quad \nabla T^*\cdot\boldsymbol{n} = 0 & \boldsymbol{x}\in\partial\mathcal{D}_2,\, t\in\mathcal{I} \\ -\lambda\nabla T^*\cdot\boldsymbol{n} = 4\epsilon\sigma T^3 T^* & \boldsymbol{x}\in\partial\mathcal{D}_3,\, t\in\mathcal{I} \end{array}\right. \tag{106}$$

and the cost gradient is written as:

$$\nabla j = -\left(T-T_\infty\right)T^*. \tag{107}$$

□

From the integration of the adjoint-state, the cost function gradient is computed. From the knowledge of the cost function gradient, the direction of descent is computed, for instance with the conjugate gradient method, or with any other faster method if a fine parameterization for h is required. It is also to be pointed out that the temperature state being varying almost linearly with the convection property, the line-search equation can be for instance given by the solution of (14).

9.3 Adjoint RTE

This last example aims at developping adjoint-state equation of the radiative transfer equation (RTE). The main objective behind this developments is the solution of optical tomography problems, in which the problem is the reconstruction of radiative properties ($\kappa(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$) within the medium, input intensity being prescribed on the boundary, and measurements being also performed on boundaries. Some of the difficulties for solving such problems include:

i) the dimension of the discrete control space is likely to be high, in 2-D and especially in 3-D. This means that efficient optimizers such as the ones based on the gradient-type BFGS are the only ones to be used. Others such as the conjugate gradients for instance may be too slow and gradient-free are not appropriate at all;

ii) the RTE is integro-differential, so appropriate inner products must be used through all the derivations. Therefore, mathematical developments for the derivation of the adjoint-state as well as for the cost function gradient must be undertaken very carefully. In the same spirit, numerical algorithm and implementation must be undertaken very carefully;

iii) the state being non linear with respect to the physical properties and overall the nonlinear inverse problem being ill-posed, several regularization strategies must be used and combined together.

Let the radiative transfer equation (RTE) being written as, $\forall\,(\boldsymbol{x},\boldsymbol{s}) \in \mathcal{D}\pi$:

$$(\boldsymbol{s}\cdot\nabla + \kappa + \sigma)\, I(\boldsymbol{x},\boldsymbol{s}) = \sigma \oint_{4\pi} I(\boldsymbol{x},\boldsymbol{s})\Phi(\boldsymbol{s},\boldsymbol{s}')\,\mathrm{d}\omega(\boldsymbol{s}'), \tag{108}$$

where $\boldsymbol{s}$ is the considered direction of propagation, $\Phi(\boldsymbol{s},\boldsymbol{s}')$ is the phase function representing the probability that a photon arriving from the direction $\boldsymbol{s}'$ is scattered to the direction $\boldsymbol{s}$, and κ and σ are the absorption and diffusion space-dependent functions, respectively. On a part of the boundary, there is a prescribed Dirichlet condition:

$$I(\boldsymbol{x},\boldsymbol{s}) = \bar{I} \quad \text{for } \boldsymbol{x} \in \partial\mathcal{D}_s \text{ and } \boldsymbol{s}\cdot\boldsymbol{n} < 0. \tag{109}$$

Also, let a cost function measuring the misfit between predictions and measurements somewhere on the boundary, i.e., $\partial\Omega_d \subset \partial\Omega$, the misfit being expressed (it is actually a norm) in terms of the radiance,

$$e = I(\boldsymbol{x},\boldsymbol{s}_d) - I_d(\boldsymbol{x},\boldsymbol{s}_d) \quad \text{for } \boldsymbol{x} \in \partial\mathcal{D}_d \text{ and } \boldsymbol{s}_d\cdot\boldsymbol{n} > 0. \tag{110}$$

In order to make the derivation of the adjoint-state, the tools described in previous sections are used. Additionally, the state variable I being defined in $(\boldsymbol{x},\boldsymbol{s}) \in \mathcal{D}\times 4\pi$, the inner product $\mathcal{U}$ defined in eq. (68) and in following equations is:

$$(u,v)_{\mathcal{U}} = \int_{4\pi}\int_{\mathcal{D}} uv\,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{s}. \tag{111}$$

After integration by parts and some – technical – manipulations in the inner products, one finds the adjoint RTE to be:

$$(-\boldsymbol{s}\cdot\nabla + \kappa + \sigma)\, I^*(\boldsymbol{x},\boldsymbol{s}) = \sigma \oint_{4\pi} I^*(\boldsymbol{x},\boldsymbol{s})\Phi(\boldsymbol{s},\boldsymbol{s}')\,\mathrm{d}\omega(\boldsymbol{s}') \tag{112}$$

coupled with the Dirichlet boundary condition:

$$I^*(\boldsymbol{x},-\boldsymbol{s}_d) = (\boldsymbol{s}_d\cdot\boldsymbol{n})^{-1}\,(I - I_d)\,(\boldsymbol{x},-\boldsymbol{s}_d) \quad \text{for } \boldsymbol{x} \in \partial\mathcal{D}_d \text{ and } \boldsymbol{s}_d\cdot\boldsymbol{n} > 0 \tag{113}$$

The adjoint-state being computed, the cost function gradient, here as an implicit function of $\boldsymbol{x}$ can be computed:

$$\begin{aligned}
\nabla_\kappa j(\boldsymbol{x}) &= \oint_{4\pi} I(\boldsymbol{x},\boldsymbol{s})I^*(\boldsymbol{x},\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} \\
\nabla_\sigma j(\boldsymbol{x}) &= \oint_{4\pi} \left(I(\boldsymbol{x},\boldsymbol{s})I^*(\boldsymbol{x},\boldsymbol{s}) - \oint_{4\pi} I(\boldsymbol{x},\boldsymbol{s}')\Phi(\boldsymbol{s},\boldsymbol{s}')\,\mathrm{d}\boldsymbol{s}'\; I^*(\boldsymbol{x},\boldsymbol{s}) \right)\,\mathrm{d}\boldsymbol{s}
\end{aligned} \tag{114}$$

This continuous version of the components of the cost function gradient is then projected onto the basis used to parameterize the control-space. A very detailed derivation of the adjoint RTE can be found in [16], for instance.

10 Concluding remarks

This lecture was devoted to the presentation of mathematical and numerical algorithms used in the estimation of functions while solving inverse heat transfer problems. Contrarily to parameter estimation problem, the dimension of the "control space" is likely to be big after the process of parameterization, this being an essential issue and reason of why using efficient optimization algorithms. Among those efficient algorithms, the ones based on the cost function gradient as well as on adjoint-states are of first importance. Functions to be estimated usually contain several regularity requirements and thus, the use of efficient regularization tools are compulsory to cope with the ill-posed character of the inverse problem.

Acknowledgements

The writing of this paper course started in 2005 for the Metti school held in Aussois, France. Since then, this course has been improved for each new Metti school edition. I would like to particularly express my gratitude to my colleagues for their invaluable help, their careful reading and reviewing as well as their suggestions of improvements:

— P. Le Masson, IRDL, Université de Bretagne Sud;
— Y. Jarny, LTEN, Université de Nantes;
— D. Maillet, LEMTA, Université de Lorraine.

References

[1] G. Allaire. *Analyse numérique et optimisation.* Les Éditions de l'École Polytechnique, Paris, Août 2005.

[2] M. Minoux. *Mathematical Programming, Theory and Applications.* Wiley, Chichester, UK, 1986.

[3] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer, New York, 1999.

[4] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization.* Academic Press, London, 1992.

[5] Y. Favennec. Optimization, sensitivity and adjoint states. "Eurotherm Winter School METTI 2005: Thermal measurements and inverse techniques: a tool for the characterization of multiphysical phenomena", Aussois, 2005.

[6] AV Goncharskii, Aleksandr Sergeevich Leonov, and Anatolii Grigor'evich Yagola. A generalized discrepancy principle. *USSR Computational Mathematics and Mathematical Physics*, 13(2):25–37, 1973.

[7] Vladimir Alekseevich Morozov, Z Nashed, and AB Aries. *Methods for solving incorrectly posed problems.* Springer, 1984.

[8] O. Alifanov. *Inverse Heat Conduction, Ill-posed problems.* Wiley Interscience, New-York, 1985.

[9] G.C. OnWubolu and B.V. Babu. *New optimization techniques in engineering.* Springer, 2003.

[10] M. Clerc. *L'optimisation par essaims particulaires.* Hermes-Lavoisier, 2005.

[11] J.C. Culioli. *Introduction à l'optimisation.* Ellipses, Paris, 1994.

[12] J. Céa. *Optimisation, théorie et algorithmes*. Dunod, Paris, 1971.

[13] B. Larrouturou and P.L. Lions. *Méthodes mathématiques pour les sciences de l'ingénieur: optimisation et analyse numérique*. Cours de l'École Polytechnique, Paris, 1994.

[14] Fabien Dubot, Yann Favennec, Benoit Rousseau, and Daniel R Rousse. Regularization opportunities for the diffuse optical tomography problem. *International Journal of Thermal Sciences*, 98:1–23, 2015.

[15] Fabien Dubot, Yann Favennec, Benoit Rousseau, and Daniel R. Rousse. A wavelet multi-scale method for the inverse problem of diffuse optical tomography. *Journal of Computational and Applied Mathematics*, 2015.

[16] Y Favennec, F Dubot, D Le Hardy, B Rousseau, and DR Rousse. Space-dependent sobolev gradients as a regularization for inverse radiative transfer problems. *Mathematical Problems in Engineering*, 2016, 2016.

Lecture 9: The Use of Techniques within the Bayesian Framework of Statistics for the Solution of Inverse Problems

Helcio R. B. Orlande[1]
[1] Department of Mechanical Engineering, Politécnica/COPPE
Federal University of Rio de Janeiro, UFRJ, Cid. Universitária
Cx. Postal: 68503, Rio de Janeiro, RJ, 21941-972, Brazil
E-mail: helcio@mecanica.coppe.ufrj.br

Abstract. In this lecture, techniques for the solution of inverse problems through statistical inference on the posterior probability density are presented. Such density is obtained through Bayes' theorem and is proportional to the product of the likelihood function, which models the measurement errors, by the prior distribution, which models the information known about the parameters before the experimental data is available. The focus of this lecture is on Markov Chain Monte Carlo (MCMC) methods. Basic concepts, as well as practical issues regarding the implementation of MCMC methods, are presented. The Metropolis-Hastings algorithm, as well as its alternative version that samples the parameters by blocks, are described in detail. Monte Carlo methods usually involve large computational times. The Approximation Error Model approach and the Delayed Acceptance Metropolis-Hastings algorithm are thus presented for computational speed up.

List of acronyms:

- **AEM:** Approximation Error Model
- **DAMH:** Delayed Acceptance Metropolis-Hastings
- **MAP**: Maximum a Posteriori
- **MCMC**: Markov Chain Monte Carlo
- **MH**: Metropolis-Hastings
- **ML**: Maximum Likelihood

Scope

1. Introduction

The term *Bayesian* is commonly used to refer to techniques for the solution of inverse problems that fall within the framework of statistics developed by the Presbyterian minister Rev. Thomas Bayes (☆1702 - †1761) [1]. Such framework was actually established after Bayes' death, when his friend, Richard Price, published Bayes' famous paper, which dealt with the following problem: "*Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between two degrees of probability that can be named*."[2]. On the other hand, it is attributed to Laplace the mathematical formulation that is known today as Bayes' theorem [3].

The term *Bayesian* was first used by R. A. Fisher, but in a pejorative context. Although born more than 120 years after the death of Bayes, Fisher was Bayes biggest intellectual rival [3]. The major issue by Fisher against Bayes and Laplace was that they used the concept of a *prior probability*, which represents the information about an unknown quantity before the measured data is available [3]. Fisher's theory relies solely on the measured data and on modelling of their associated uncertainty, aiming at unbiased inference and/or decision; therefore, it is usually referred to as the *frequentist* framework for statistics [1,3,4]. On the other hand, within the Bayesian framework credit is also given to previous information, in addition to that given to the measured data. Such previous information can even be qualitative, but needs to be represented in terms of a probability distribution function, and regretfully induces bias in the results [1,3,4]. Nevertheless, the use of prior information in the Bayesian framework does not mean that it completely overtakes the information provided by the measured data, unless the last one is too uncertain to be really taken into account. Interestingly, one may also argue that life is Bayesian: think about life as a sequential process and notice that everyday our past beliefs are combined with new observable data, in order to provide a better understanding about different matters of our interest, like the faster way to go to work, people, natural phenomena, industrial processes (and their effects on nature, like climate change), etc.

Although not always considered in such a way, the solution of inverse problems can be appropriately formulated in terms of statistical inference [5]. Statistical inference refers to the process of drawing conclusions or making predictions based on limited information, beyond the immediate data that is available [4]. Note that this is exactly what is aimed with the solution of inverse problems, which can be broadly defined as those dealing with the estimation of unknown quantities appearing in mathematical models, by using measurements of some of their dependent variables (observable response of the problem) and their computational solution (estimated response of the problem) [5-27].

There are many techniques for the solution of inverse problems, but the most general ones are usually related to the minimization of an objective function that involves the difference between measured and estimated responses of a problem [5-27]. If the objective function is derived based on statistical hypotheses for the measurement errors and for the unknown parameters/functions, the minimization procedure can be related to statistical inference, thus resulting in point estimates for the unknowns that allow for estimations of their associated uncertainties [5,8]. Unfortunately, such is generally not the case, in special when the objective function is penalized with regularization terms.

The solution of inverse problems within the Bayesian framework of statistics is recast in the form of inference on the so-called *posterior probability density*, which is the model for the conditional probability distribution of the unknown parameters given the measurements. The measurement model incorporating the related uncertainties is called the *likelihood*, that is, the conditional probability of the measurements given the unknown parameters. The model for the unknowns that reflects all the uncertainty of the parameters without the information conveyed by the measurements, is called the *prior* distribution [5,8,20,22,25-29]. The prior distribution is combined with the likelihood to form the posterior distribution by using Bayes' theorem [5,8,20,22,25-29].

The objective of this text is to introduce some basic concepts regarding the solution of inverse heat transfer problems within the Bayesian framework of statistics. Emphasis is given to the use of *Markov Chain Monte Carlo (MCMC) methods* [1,4,5,20,22,25-29]. Monte Carlo methods are also designated as *stochastic simulation techniques*, since values simulated (sampled) from the distribution of interest, which in general is not completely known, are used for the computation of its statistics [28]. Simulation techniques rely on probabilistic results, such as the law of large numbers and the central limit theorem, which ensure that the approximate statistics tend to the actual ones as the number of simulations increase [28].

This text is not aimed at a literature review about the subject, which would certainly include a very large number of works ranging from statistical, mathematical and computational aspects, to practical engineering applications. Indeed, an analysis of recent conferences on inverse problems clearly shows a trend of increasing number of papers that make use of solution techniques within the Bayesian framework of statistics, as faster computers become available. This text also does not cover Bayesian filters for the solution of state estimation problems.

It is the author's opinion that the most complete source for the solution of inverse problems within the Bayesian framework of statistics is the book by Kaipio and Somersalo [5]. The reader is referred to the book by Gamerman and Lopes [28] and to the book edited by Brooks et al. [29] for deeper details about Markov Chain Monte Carlo methods. Fundamental material on Bayesian statistics can be found in the books by Lee [1] and Winkler [4]. A very didactical series of videos presenting Monte Carlo Markov Chain methods can be found at https://www.youtube.com/watch?v=12eZWG0Z5gY. Two interesting books, with historical aspects and practical applications of Bayesian statistics in layman's terms, include references [3] and [30]. It is highly recommended that the reader consults the texts of Tutorials 9 and 14 of this METTI School, for the implementation and computational speed-up of Markov Chain Monte Carlo methods.

2. General Considerations

Consider the mathematical formulation of a heat transfer problem, which, for instance, can be linear or non-linear, one or multi-dimensional, involve one single or coupled heat transfer modes, etc. We denote the vector of parameters appearing in such formulation as:

$$\mathbf{P}^T = \left[P_1, P_2, ..., P_N \right] \tag{1}$$

where N is the number of parameters. These parameters can possibly be thermal conductivity components, heat transfer coefficients, heat sources, boundary heat fluxes, etc. They can represent constant values of such quantities, or parameters in the representation of a function in terms of known basis functions. For example, we can consider a transient heat source term $g_p(t)$ parameterized as follows:

$$g_p(t) = \sum_{j=1}^{N} P_j C_j(t) \tag{2.a}$$

where $C_j(t)$, $j = 1,\ldots,N$, are linearly-independent basis functions that generate the space of the projected function $g_p(t)$ onto a space of finite dimension N. Note that $C_j(t)$ can also be functions with local support, such as,

$$C_j(t) = \begin{cases} 1 \quad , \quad \text{for } t_j - \dfrac{t_j - t_{j-1}}{2} < t < t_j + \dfrac{t_{j+1} - t_j}{2} \\ 0 \quad , \quad \text{elsewhere} \end{cases} \tag{2.b}$$

With the basis functions given by Eq. (2.b), each parameter P_j represents the local value of the function in the time interval $t_j - \dfrac{t_j - t_{j-1}}{2} < t < t_j + \dfrac{t_{j+1} - t_j}{2}$, that is, $g_p(t_j) = P_j$ as illustrated by Figure 1.

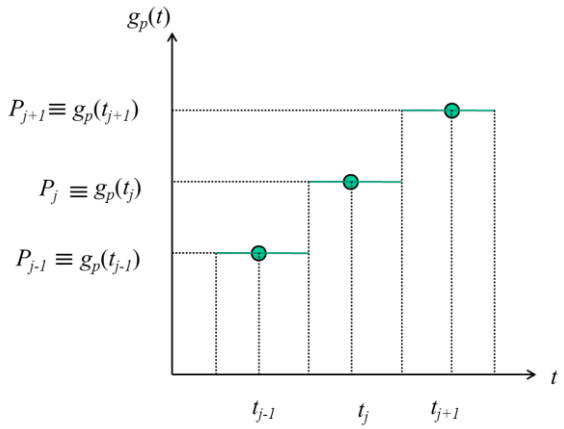


Figure 1. Parameters representing local values of a function that varies in time.

Consider also that transient measurements are available regarding the heat transfer processes being mathematically formulated. The vector containing the measurements is written as:

$$\mathbf{Y}^T = \left(\vec{Y}_1 \ , \ \vec{Y}_2 \ , \ldots , \vec{Y}_I \right) \tag{3.a}$$

where $\vec{Y}_i$ contains the data of M sensors at time t_i, $i = 1, \ldots, I$, that is,

$$\vec{Y}_i = \left(Y_{i1} \ , \ Y_{i2} \ , \ldots , Y_{iM} \right) \quad \text{for } i=1,\ldots,I \tag{3.b}$$

Therefore, we have $D = M\,I$ measurements in total. Note that, in practice, the measured data are not limited to temperatures, but could also include heat fluxes, radiation intensities, etc.

Throughout this tutorial, the measurement errors, $\boldsymbol{\varepsilon}$, are assumed to be additive, that is,

$$\mathbf{Y} = \mathbf{T}(\mathbf{P}) + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{T}(\mathbf{P})$ is a highly accurate computational solution of the mathematical formulation of the problem under analysis, obtained with the vector of parameters $\mathbf{P}$, that is,

$$\mathbf{T}^T(\mathbf{P}) = [\,\vec{T}_1(\mathbf{P}) \ , \quad \vec{T}_2(\mathbf{P}) \ , \ldots , \quad \vec{T}_I(\mathbf{P})\,] \tag{5.a}$$

where

$$\vec{T}_i(\mathbf{P}) = [\,T_{i1}(\mathbf{P}) \ , \ T_{i2}(\mathbf{P}) \ , \ldots , \ T_{iM}(\mathbf{P})] \quad \text{for } i=1,\ldots,I \tag{5.b}$$

The mathematical formulation in $\mathbf{T}(\mathbf{P})$ is supposed to represent the problem of interest, that is, the experimental data $\mathbf{Y}$, with the least possible amount of uncertainty. It is thus referred to as a *high-fidelity* model. Anyhow, approximation errors resulting from the replacement of the high-fidelity model by a low-fidelity model for the solution of the inverse problem can be formally taken into account within the Bayesian framework of statistics [5]. An approach to deal with approximation errors will be described later in this text, and the reader should also refer to Tutorial 14 of this METTI School.

By further assuming that the measurement errors, $\boldsymbol{\varepsilon}$, are Gaussian random variables, with zero means, known covariance matrix $\mathbf{W}$ and independent of the parameters $\mathbf{P}$, their probability density function, $p(\boldsymbol{\varepsilon})$, is given by [5,8,20,22,25-29]:

$$p(\boldsymbol{\varepsilon}) = (2\pi)^{-D/2} \left|\mathbf{W}\right|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{\varepsilon}^T \mathbf{W}^{-1} \boldsymbol{\varepsilon} \right\} \tag{6.a}$$

Due to the additive model for the measurement errors given by equation (4), equation (6.a) can be rewritten as

$$p(\boldsymbol{\varepsilon}) = (2\pi)^{-D/2} \left|\mathbf{W}\right|^{-1/2} \exp\left\{ -\frac{1}{2} [\mathbf{Y} - \mathbf{T}(\mathbf{P})]^T \mathbf{W}^{-1} [\mathbf{Y} - \mathbf{T}(\mathbf{P})] \right\} \tag{6.b}$$

Thus, $p(\boldsymbol{\varepsilon}) = p(\mathbf{Y} \,|\, \mathbf{P})$, which is the conditional probability density of different measurement outcomes $\mathbf{Y}$ with a fixed $\mathbf{P}$, denoted as the *likelihood function* [5,8,20,22,25-29].

A very common approach for the solution of inverse problems, dealing with the estimation of the parameters $\mathbf{P}$ by using the measurements $\mathbf{Y}$, is to maximize the likelihood function. With

the above hypotheses regarding the measurement errors, this can be accomplished through the minimization of the absolute value of the term inside the exponential function of equation (6.b), resulting in the following *maximum likelihood objective function*:

$$S_{ML}(\mathbf{P}) = \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right]^T \mathbf{W}^{-1} \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right] \tag{7}$$

The least squares norm can be obtained as a particular case of Eq. (7), if the measurements are also considered as uncorrelated and with constant variances σ^2, in addition to the above hypotheses [8]. In this case, the covariance matrix $\mathbf{W}$ is given by:

$$\mathbf{W} = \sigma^2 \mathbf{I} \tag{8}$$

where $\mathbf{I}$ is the identity matrix. Then, the minimization of Eq. (7) is equivalent to the minimization of the least squares norm:

$$S_{OLS}(\mathbf{P}) = \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right]^T \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right] \tag{9}$$

The covariance matrix of the values estimated for the parameters $\mathbf{P}$ with the minimization of equation (7), is given by [8]:

$$\text{cov}(\mathbf{P}) = (\mathbf{J}^T \mathbf{W}^{-1} \mathbf{J})^{-1} \tag{10.a}$$

which reduces to:

$$\text{cov}(\mathbf{P}) = (\mathbf{J}^T \mathbf{J})^{-1} \sigma^2 \tag{10.b}$$

if $\mathbf{W}$ is given by equation (8). Equations (10.a,b) are exact for linear estimation problems, but can be used as approximations for nonlinear problems [8].

Therefore, in order to make use of the minimization of the least squares norm for obtaining point estimates for the parameters $\mathbf{P}$ that have some statistical meaning (for example, that allow estimates of the covariances of the estimated parameters with equation 10.b), all the statistical hypotheses stated above need to be valid [8]. Such a fact is quite often overlooked when an objective function is defined for the solution of an inverse problem via optimization techniques. Still, if the estimation problem is linear, the measurement errors are additive, with zero mean, and with a covariance matrix that is positive definite and known to within a multiplicative constant σ^2, that is,

$$\mathbf{W} = \hat{\mathbf{W}} \sigma^2 \tag{11}$$

the Gauss-Markov theorem [8,18] assures that minimum variance estimates can be obtained with the minimization of:

$$S_{GM}(\mathbf{P}) = \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right]^T \hat{\mathbf{W}}^{-1} \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right] \tag{12}$$

even if the measurement errors are not Gaussian. In such a case, if $\hat{\mathbf{W}} = \mathbf{I}$ the minimization of the ordinary least squares norm provides minimum variance estimates. On the other hand, the covariance matrix of the values estimated for the parameters $\mathbf{P}$ cannot be computed with equations (10.a,b) since σ^2 is not known.

Different methods can be used for the minimization of equations (7), (9) or (12), after an analysis of the sensitivity coefficients of the parameters and an appropriate experimental design [8-26]. For a linear case, the minimization of equation (7) is obtained with:

$$\mathbf{P} = (\mathbf{J}^T \mathbf{W}^{-1} \mathbf{J})^{-1} \mathbf{J}^T \mathbf{W}^{-1} \mathbf{Y} \tag{13.a}$$

while, for the nonlinear case the iterative procedure of Gauss' method gives:

$$\mathbf{P}^{k+1} = \mathbf{P}^k + (\mathbf{J}^T \mathbf{W}^{-1} \mathbf{J})^{-1} \mathbf{J}^T \mathbf{W}^{-1} [\mathbf{Y} - \mathbf{T}(\mathbf{P}^k)] \tag{13.b}$$

where the superscript k denotes the number of iterations and $\mathbf{J}$ is the sensitivity matrix.

We note that other *maximum a posteriori* objective functions can be derived if the measurement errors follow density functions different from the Gaussian distribution examined above.

3. Bayesian Framework of Statistics

For the solution of inverse problems within the Bayesian framework of statistics, all variables included in the mathematical formulation of the physical problem are modelled as random variables. Techniques for the solution of inverse problems within the Bayesian framework of statistics can be summarized by the following steps [5]:

1. Based on all information available for the parameters $\mathbf{P}$ before the measured data $\mathbf{Y}$ is taken, select a probability distribution function, $p(\mathbf{P})$, that appropriately represents the *prior* information.
2. Select the *likelihood function*, $p(\mathbf{Y}|\mathbf{P})$, that appropriately models the measurement errors. The likelihood function involves a relation between the experimental observations and the computed solutions of the high-fidelity mathematical model of the problem under analysis (see, for example, equation 6.b).
3. Develop methods to explore the *posterior* density function, which is the conditional probability distribution of the unknown parameters given the measurements, $p(\mathbf{P}|\mathbf{Y})$.

The formal mechanism to combine the new information (measurements) with the previously available information (prior) is known as Bayes' theorem [5,8,20,22,25-29]. Let $\mathbf{P}$ and $\mathbf{Y}$ be continuous random variables. Then, we can write [4]:

$$p(\mathbf{P}|\mathbf{Y}) = \frac{p(\mathbf{P}, \mathbf{Y})}{p(\mathbf{Y})} \tag{14}$$

that is, the conditional density of the random variable $\mathbf{P}$ given a value of the random variable $\mathbf{Y}$ is the joint density of $\mathbf{P}$ and $\mathbf{Y}$ divided by the marginal density of $\mathbf{Y}$, where:

$$p(\mathbf{Y}) = \int_{\square^N} p(\mathbf{P}, \mathbf{Y})\, d\mathbf{P} \tag{15}$$

The joint density $p(\mathbf{P}, \mathbf{Y})$ is not generally known, but it can be written in terms of the likelihood and the prior as [4]:

$$p(\mathbf{P}, \mathbf{Y}) = p(\mathbf{Y}|\mathbf{P})\, p(\mathbf{P}) \tag{16}$$

By substituting (16) into (14) we then obtain Bayes' theorem, which is given by:

$$p(\mathbf{P}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{P})\; p(\mathbf{P})}{p(\mathbf{Y})} \tag{17.a}$$

where $p_{posterior}(\mathbf{P}) = p(\mathbf{P}|\mathbf{Y})$ is the posterior probability density, $p(\mathbf{P})$ is the prior density, $p(\mathbf{Y}|\mathbf{P})$ is the likelihood function and $p(\mathbf{Y})$ is the marginal probability density of the measurements (also called evidence), which plays the role of a normalizing constant. Since the computation of $p(\mathbf{Y})$ with equation (15) is in general difficult, and usually not needed for practical calculations as will be apparent below, Bayes' theorem is commonly written as:

$$p_{posterior}(\mathbf{P}) = p(\mathbf{P}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{P})\, p(\mathbf{P}) \tag{17.b}$$

4. Maximum a Posteriori Objective Function

Consider a case with a Gaussian prior density model for the unknown parameters in the form:

$$p(\mathbf{P}) = (2\pi)^{-N/2} \left|\mathbf{V}\right|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{P} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{P} - \boldsymbol{\mu}) \right] \tag{18}$$

where $\boldsymbol{\mu}$ and $\mathbf{V}$ are the known mean and covariance matrix for $\mathbf{P}$, respectively. By assuming normally distributed measurement errors with zero means and known covariance matrix $\mathbf{W}$, which are also supposed additive and independent of the parameters $\mathbf{P}$, the likelihood function is given by equation (6.b). By substituting equations (6.b) and (18) into Bayes' theorem given by equation (17.b), we obtain:

$$\ln\left[p(\mathbf{P} \mid \mathbf{Y}) \right] \propto -\frac{1}{2} \left[(D + N) \ln 2\pi + \ln |\, \mathbf{W} \,| + \ln |\, \mathbf{V} \,| + S_{MAP}(\mathbf{P}) \right] \tag{19}$$

where

$$S_{MAP}(\mathbf{P}) = \left[\mathbf{Y} - \mathbf{T}(\mathbf{P}) \right]^T \mathbf{W}^{-1} \left[\mathbf{Y} - \mathbf{T}(\mathbf{P}) \right] + (\mathbf{P} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{P} - \boldsymbol{\mu}) \tag{20}$$

Equation (19) reveals that the maximization of the posterior distribution can be obtained with the minimization of the objective function given by equation (20), denoted as the *maximum a posteriori* (MAP) *objective function* for the statistical hypotheses made above [5,8,20,22,25-29]. Equation (20) shows the contributions of the likelihood and of the prior distributions in this objective function, given by the first and second terms on the right-hand side, respectively. It is now interesting to notice that the maximum likelihood objective function (equation 7) is not a Bayesian estimator, since it does not contain information provided by the prior distribution for the parameters. Conspicuously, the least squares norm (equation 9) and other objective functions derived from equation (7), even those containing penalization terms (e.g., Tikhonov's regularization), are not Bayesian estimators, since they only explore the information provided by the measurements and, eventually, some characteristics of the parameters, like smoothness. Although the second term on the right-hand side of equation (20) is a quadratic form and resembles Tikhonov's regularization, there is a fundamental difference between the two approaches. Tikhonov's regularization focuses on obtaining a stabilized form of the original objective function and is not designed to yield uncertainty estimates that would have a statistical interpretation. In contrast, Bayesian inference assumes that the likelihood and prior statistical models reflect their actual uncertainties. Hence, uncertainties computed from equation (19) correspond to the actual posterior uncertainties only if this hypothesis is valid [5].

Such as for the maximum likelihood objective function, different methods can be used for the minimization of equation (20) in order to obtain point estimates for the unknowns. For nonlinear problems, the Gauss method results in the following iterative procedure [5,8,20,22,25-29]:

$$\mathbf{P}^{k+1} = \mathbf{P}^{k} + [\mathbf{J}^{T}\mathbf{W}^{-1}\mathbf{J} + \mathbf{V}^{-1}]^{-1}\{\mathbf{J}^{T}\mathbf{W}^{-1}[\mathbf{Y} - \mathbf{T}(\mathbf{P}^{k})] + \mathbf{V}^{-1}(\boldsymbol{\mu} - \mathbf{P}^{k})\} \tag{21}$$

Note in equation (21) that the conditioning of the matrix $\mathbf{J}^{T}\mathbf{W}^{-1}\mathbf{J}$ can be improved with the matrix $\mathbf{V}^{-1}$, which is the inverse of the covariance matrix of the Gaussian prior for the parameters. Therefore, the estimation of the parameters can be stabilized by using prior information with small covariances. Despite such desired effect for the regularization of the estimation procedure, the MAP estimator is biased and the expected value of $\mathbf{P}$ is $\boldsymbol{\mu}$ [8]. Such a fact clearly shows the important requirement of modeling the prior information as accurately as possible for the success of the inverse analysis within the Bayesian framework of statistics. For a linear case, the covariance matrix of the posterior Gaussian distribution is given by [8]:

$$\mathrm{cov}(\mathbf{P}) = (\mathbf{J}^{T}\mathbf{W}^{-1}\mathbf{J} + \mathbf{V}^{-1})^{-1} \tag{22}$$

which can be used as an approximation for nonlinear cases.

A comparison of equation (22) with the covariance matrix related to the maximum likelihood objective function (equation 10.a) shows that the covariance of the prior, $\mathbf{V}$, is reduced by solving the inverse problem if $\mathbf{J}^{T}\mathbf{W}^{-1}\mathbf{J}$ is well conditioned. Therefore, the solution of the inverse problem improves the information *a priori* available for the parameters, if the sensitivity coefficients are linearly independent and with large magnitudes, that is, the determinant of $\mathbf{J}^{T}\mathbf{J}$ is large and $\mathbf{J}^{T}\mathbf{W}^{-1}\mathbf{J}$ is well conditioned.

5. Markov Chain Monte Carlo (MCMC) Methods

The Gaussian likelihood and the Gaussian prior examined in section 4 resulted in an expression for the posterior (equation 19) from which a MAP point estimate can be obtained for the parameters, provided that the minimum of equation (20) exists. In this particular case (Gaussian likelihood and Gaussian prior), the prior is *conjugate* to the likelihood [1,4,5,28]. A class Π of prior distributions is said to form a conjugate family if the posterior density is in the same class Π for all $\mathbf{P}$, whenever the prior density is in Π [1]. Although this property is valid for many cases that involve continuous distributions, in special those that belong to the exponential family [1,28], the posterior probability distribution may not allow an analytical treatment if non-conjugate prior probability densities are assumed for the parameters. Moreover, whereas the computation of the MAP estimate is an optimization problem, that is,

$$\mathbf{P}_{MAP} = \arg\max_{\mathbf{P}\in\Omega^N} p(\mathbf{P}\,|\,\mathbf{Y}) \tag{23}$$

other point and confidence estimates from the posterior distribution typically require numerical integration. For example, one common point estimate is the conditional mean defined as [5]:

$$\mathbf{P}_{CM} = E(\mathbf{P}) = \int_{\Omega^N} \mathbf{P}\; p(\mathbf{P}\,|\,\mathbf{Y})\, d\mathbf{P} \tag{24}$$

where $E(.)$ denotes the expected value. In general, the dimension N of the parameter space is large enough to make the numerical integration in equation (24) impractical. Besides that, the computation of the normalizing constant in the denominator of $p(\mathbf{P}\,|\,\mathbf{Y})$ (see equations 14-17) already constitutes a challenging problem by itself.

For those cases that the posterior is not analytical and/or numerical integrations required for estimates are not practical, Markov Chain Monte Carlo (MCMC) methods can provide a solution of the inverse problem, so that inference on the posterior probability density becomes inference on its samples [1,4,5,20,22,25-28]. For example, the Monte Carlo integration of equation (24) can be approximated by [5]:

$$\mathbf{P}_{CM} = E(\mathbf{P}) = \int_{\Omega^N} \mathbf{P}\; p(\mathbf{P}\,|\,\mathbf{Y})\, d\mathbf{P} \approx \frac{1}{n}\sum_{t=1}^{n} \mathbf{P}^{(t)} \tag{25}$$

where $\mathbf{P}^{(t)}$, for $t = 1,\ldots, n$, are samples from $p(\mathbf{P}\,|\,\mathbf{Y})$. Markov Chain Monte Carlo methods are used to obtain such samples.

Due to the simplicity in the application of MCMC methods, such a technique for the solution of inverse problems has been recently becoming quite popular, being applied even for cases where a MAP estimate would be feasible. One clear disadvantage on the application of Monte Carlo methods is the required large computational times. On the other hand, the use of computationally fast low-fidelity models can be appropriately accommodated within the Bayesian framework of statistics [5], so that the application of MCMC methods to many practical problems is nowadays possible.

Concepts and properties of Markov chains are presented in this section, which is then finished with a powerful, simple and popular MCMC algorithm. Some practical aspects and speedup techniques for the implementation of MCMC methods are delayed to other sections further below.

5.1. Markov Chains

Markov chains are named after the Russian mathematician A. A. Markov, who developed such concept by investigating the alternance of vowels and consonants in a Russian poem. Poincaré also dealt with sequences of random variables that were in fact Markov chains [28]. A Markov chain is a stochastic process that, given the present state, past and future states are independent. The collection of the random quantities $\{\mathbf{P}^{(t)} : t \in T\}$ is said to be a stochastic process with state space S and index set T. The state space is a subset of $\mathbb{R}^N$, that is, the support of the parameter vector $\mathbf{P}$, while here T is the set of Natural numbers that index the states of the Markov chain [28].

The stochastic process is a Markov chain if it satisfies the Markov condition [1,4,5,20,22,25-29]:

$$q(\mathbf{P}^{(t+1)} = \mathbf{y} \mid \mathbf{P}^{(t)} = \mathbf{x}, \mathbf{P}^{(t-1)} = \mathbf{x}^{(t-1)}, \ldots, \mathbf{P}^{(0)} = \mathbf{x}^{(0)}) = q(\mathbf{P}^{(t+1)} = \mathbf{y} \mid \mathbf{P}^{(t)} = \mathbf{x})$$
$$\text{for all} \quad \mathbf{y}, \mathbf{x}, \mathbf{x}^{(t-1)}, \ldots, \mathbf{x}^{(0)} \in S \tag{26}$$

where q is a transition probability. Some concepts regarding Markov chains are now presented. The reader shall consult references [1,4,5,20,22,25-29] for further details.

If the transition probability does not depend on t, that is, if:

$$q(\mathbf{P}^{(t+m+1)} = \mathbf{y} \mid \mathbf{P}^{(t+m)} = \mathbf{x}) = q(\mathbf{P}^{(t+1)} = \mathbf{y} \mid \mathbf{P}^{(t)} = \mathbf{x}) \qquad \text{for all } m \in T \tag{27}$$

the Markov chain is said to be *homogenous* [22].

A distribution p^* is said to be a *stationary distribution* of a chain if, once the chain is in p^* it stays in this distribution. Suppose now that $p^{(t)} \to p^*$ as $t \to \infty$ for any $p^{(0)}$, where $p^{(t)}$ is the distribution at state t of the chain. Then, p^* is the *equilibrium distribution* of the Markov chain and the chain is said to be *ergodic*.

Consider the sequence of states $\mathbf{x} \to \mathbf{P}^{(1)} \to \mathbf{P}^{(2)} \to \cdots \mathbf{P}^{(t)} \to \mathbf{y}$ so that the transition probabilities $q(\mathbf{P}^{(1)} \mid \mathbf{x}) \neq 0$, $q(\mathbf{P}^{(2)} \mid \mathbf{P}^{(1)}) \neq 0$, … , $q(\mathbf{y} \mid \mathbf{P}^{(t)}) \neq 0$. Then, there is a sequence of states from $\mathbf{x}$ to $\mathbf{y}$ with a nonzero probability of occurring in the Markov chain. It is said that $\mathbf{x}$ and $\mathbf{y}$ communicate. If $\mathbf{y}$ and $\mathbf{x}$ also communicate through nonzero transition probabilities, it is said that these two states intercommunicate. If all states in S intercommunicate, then the state space is said to be *irreducible* under q. A Markov chain is *reversible* if $p(\mathbf{x})\, q(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y})\, q(\mathbf{x} \mid \mathbf{y})$.

The period of a state $\mathbf{x}$, denoted by d_x, is the largest common divisor of the set $\{m \geq 1 : q^{(m)}(\mathbf{x},\mathbf{x}) > 0\}$. A state $\mathbf{x}$ is aperiodic if $d_x = 1$. A chain is *aperiodic* if all its states are aperiodic.

5.2. Metropolis-Hastings Algorithm

The most common MCMC algorithms are the Gibbs Sampler and the Metropolis-Hastings algorithm [1,4,5,20,22,25-29]. The Gibbs Sampler is not presented here for the sake of brevity. The Metropolis-Hastings algorithm was first devised by Metropolis et al. [31] in 1953, who aimed at the calculation of the properties of substances composed of interacting molecules. It was, therefore, a work focused on statistical mechanics, not in statistics (or inverse problems!) itself. Although the paper has five co-authors [31], only the name of the first author became popular to designate the developed algorithm, which was lately generalized by Hastings in 1970 [32]. In fact, there are some controversies about who actually contributed on the work by Metropolis et al. [33].

The above concepts about Markov chains allow the statement of the following result regarding the Metropolis-Hastings algorithm [22]: *Let p be a given probability distribution. The Markov chain simulated by the Metropolis-Hastings algorithm is reversible with respect to p. If it is also irreducible and aperiodic, then it defines an ergodic Markov chain with unique equilibrium distribution p^*.*

Unfortunately, it might not be possible to prove that the chain is irreducible and/or aperiodic for practical cases. In fact, parameters with linearly dependent sensitivity coefficients generally result on periodic and correlated chains and an equilibrium distribution is not reached. Similar to classical methods of parameter estimation, where the sensitivity coefficients directly influence the topology of the objective function based on the likelihood (see equation 7, for example) and a global minimum might not exist, such coefficients directly influence the posterior distribution, which is now sought via the implementation of a Markov chain. Therefore, the sensitivity coefficients need also to be carefully examined if the solution of the inverse parameter estimation problem is to be obtained within the Bayesian framework of statistics. In classical methods based on the maximum likelihood objective function, parameters with small and linearly dependent sensitivity coefficients are usually deterministically fixed, based on values known from previous experience and/or literature. In approaches within the Bayesian framework of statistics, uncertainties on such kind of parameters can be appropriately taken into account through their prior distribution functions. However, parameters with small and/or linearly dependent sensitivity coefficients require informative prior distributions for the success of the estimation procedure.

The Metropolis-Hastings algorithm draws samples from a candidate density by following acceptance-rejection sampling [1]. The acceptance-rejection method is used to generate samples from a density $p(\mathbf{P}) = \tilde{p}(\mathbf{P}) / K$, where the normalizing constant K might be unknown, such as in the posterior distribution given by Bayes' theorem (equation 17.a). Instead of sampling from $p(\mathbf{P})$, assume that there exists a candidate density $h(\mathbf{P})$ that is easy to simulate samples from, where $\tilde{p}(\mathbf{P}) \leq c\, h(\mathbf{P})$ and c is a known constant. The following steps

are then used to obtain a random variable $\hat{\mathbf{P}}$ from density $p(\mathbf{P})$ with the acceptance-rejection method [1]:

1. Generate a random variable $\mathbf{P}^*$ from the density $h(\mathbf{P})$;
2. Generate a random value $U \sim \mathrm{U}(0,1)$, which is uniformly distributed in (0,1);
3. If $U \leq \dfrac{\tilde{p}(\mathbf{P}^*)}{c\,h(\mathbf{P}^*)}$, let $\hat{\mathbf{P}} = \mathbf{P}^*$. Otherwise, return to step 1.

The implementation of the Metropolis-Hastings algorithm starts with the selection of a candidate or proposal distribution $q(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, which is used to draw a new candidate sample $\mathbf{P}^*$ given the current sample $\mathbf{P}^{(t)}$ of the Markov chain. For the solution of the inverse problem within the Bayesian framework of statistics, one aims at simulating the posterior distribution $p_{posterior}(\mathbf{P}) \propto p(\mathbf{Y}|\mathbf{P})\,p(\mathbf{P})$ (see equation 17.b). Hence, the balance (reversibility) condition of the Markov chain of interest is given by:

$$p_{posterior}(\mathbf{P}^{(t)})\,q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = p_{posterior}(\mathbf{P}^*)\,q(\mathbf{P}^{(t)} \mid \mathbf{P}^*) \tag{28}$$

In order to avoid eventual cases that $p_{posterior}(\mathbf{P}^{(t)})\,q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) > p_{posterior}(\mathbf{P}^*)\,q(\mathbf{P}^{(t)} \mid \mathbf{P}^*)$, that is, the process moves from $\mathbf{P}^{(t)}$ to $\mathbf{P}^*$ more often than the reverse, a probability $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ is introduced in equation (28), so that [1]:

$$p_{posterior}(\mathbf{P}^{(t)})\,q(\mathbf{P}^* \mid \mathbf{P}^{(t)})\,\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = p_{posterior}(\mathbf{P}^*)\,q(\mathbf{P}^{(t)} \mid \mathbf{P}^*) \tag{29}$$

Therefore,

$$\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p_{posterior}(\mathbf{P}^*)\,q(\mathbf{P}^{(t)} \mid \mathbf{P}^*)}{p_{posterior}(\mathbf{P}^{(t)})\,q(\mathbf{P}^* \mid \mathbf{P}^{(t)})}\right] \tag{30}$$

where $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = 1$ when the balance condition is satisfied.

Equation (30) is also called the Metropolis-Hastings ratio. Notice that there is no need to know the normalizing constant that appears in the definition of the posterior distribution (see equations 17.a,b) for the computation of equation (30). Equation (29) shows that the probability of moving from the sample at the current state $\mathbf{P}^{(t)}$ to $\mathbf{P}^*$ is now given by $[q(\mathbf{P}^* \mid \mathbf{P}^{(t)})\,\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})]$.

In the Metropolis-Hastings algorithm, a candidate $\mathbf{P}^*$ is accepted, such as in the acceptance-rejection method described above. The Metropolis-Hastings algorithm can then be summarized in the following steps [1,4,5,20,22,25-29]:

1. Let $t = 0$ and start the Markov chain with sample $\mathbf{P}^{(0)}$ at the initial state.
2. Sample a candidate point $\mathbf{P}^*$ from a proposal distribution $q(\mathbf{P}^* \mid \mathbf{P}^{(t)})$.
3. Calculate the probability $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ with equation (30).
4. Generate a random value $U \sim \mathrm{U}(0,1)$, which is uniformly distributed in (0,1).
5. If $U \leq \alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, set $\mathbf{P}^{(t+1)} = \mathbf{P}^*$. Otherwise, set $\mathbf{P}^{(t+1)} = \mathbf{P}^{(t)}$.
6. Make $t = t+1$ and return to step 2 in order to generate the sequence $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(n)}\}$.

In this way, a sequence is generated to represent the posterior distribution and inference on this distribution is obtained from inference on the samples $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(n)}\}$. We note that values of $\mathbf{P}^{(t)}$ must be ignored from $t = 0$ until the chain has not converged to equilibrium (the burn-in period).

For the computational implementation of the Metropolis-Hastings algorithm, the test in step 5 is performed by taking the logarithm of both sides, that is, $\ln[U] \leq \ln[\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})]$. This is required in order to avoid numerical errors, since $p_{posterior}(\mathbf{P})$ commonly involve exponentials and the ratio in $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ may become a number that cannot be represented within the computer numerical limits if $p_{posterior}(\mathbf{P}^{(t)})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) << p_{posterior}(\mathbf{P}^*)\, q(\mathbf{P}^{(t)} \mid \mathbf{P}^*)$.

5.3. Proposal Distributions

The proposal distribution plays a fundamental role for the success of the Metropolis-Hastings algorithm. Typical choices for $q(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ are presented below.

(i) Random Walk: In this case $\mathbf{P}^* = \mathbf{P}^{(t)} + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a vector of random variables with distribution $q_1(\boldsymbol{\psi})$. Therefore, $q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = q_1(\boldsymbol{\Psi})$. If the proposal distribution is symmetric, that is, $q_1(\boldsymbol{\psi}) = q_1(-\boldsymbol{\psi})$ or $q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = q(\mathbf{P}^{(t)} \mid \mathbf{P}^*)$, equation (30) reduces to

$$\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p_{posterior}(\mathbf{P}^*)}{p_{posterior}(\mathbf{P}^{(t)})}\right] \tag{31}$$

Thus, for this choice of the proposal density, equation (31) shows that in step 5 of the Metropolis-Hastings algorithm the candidate point $\mathbf{P}^*$ is always accepted if the move leads to a region of larger posterior probability. Furthermore, the candidate point can also be accepted if $p_{posterior}(\mathbf{P}^*) < p_{posterior}(\mathbf{P}^{(t)})$ with probability $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, thus allowing that the state space be highly explored.

Uniform and Gaussian distributions are commonly used for $q_1(\boldsymbol{\psi})$. Consider one single component P_j (j = 1,..., N) of the vector $\mathbf{P}$. For a random walk proposal with a uniform distribution one can write:

$$P_j^* = P_j^{(t)} + w_j(2u_j - 1) \tag{32.a}$$

where u_j is a random number with uniform distribution in (0,1), that is, $u_j \sim \mathrm{U}(0,1)$, while w_j is the maximum variation to generate the candidate parameter at each state of the Markov chain.

For a random walk proposal with a Gaussian distribution, we have:

$$P_j^* = P_j^{(t)} + r_j \tag{32.b}$$

where now r_j is a Gaussian random number with zero mean and standard deviation ξ_j.

The probability of accepting the candidate P_j^* increases with small variations w_j or with small standard deviations ξ_j, j = 1,..., N, in the random walk proposal with uniform and Gaussian distributions, respectively. Such is the case because it is more likely to move to regions of higher posterior around $P_j^{(t)}$ with small w_j or ξ_j. With candidates generated from small perturbations of $P_j^{(t)}$, the number of accepted states can thus be large and the resulting Markov chains may take too long to reach an equilibrium distribution for the parameters. On the other hand, large perturbations of $P_j^{(t)}$ may lead to small acceptance rates, meaning that the parameter values at the current state may be repeated at many successive states in the Markov chain, in accordance with step 5 of the Metropolis-Hastings algorithm. Although large perturbations of $P_j^{(t)}$ can fast lead to an equilibrium distribution, long chains may still be needed to generate enough samples with different (and independent) values that can be used to represent the posterior distribution of the parameters.

(ii) Independent Move: This choice for the proposal density is of the kind $q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = q_2(\mathbf{P}^*)$, that is, it does not depend on the current state $\mathbf{P}^{(t)}$. In this case, the proposal density $q(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ can be conveniently selected as the prior density $p(\mathbf{P}^*)$. By utilizing equation (17.b), equation (30) is rewritten as

$$\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p(\mathbf{Y} \mid \mathbf{P}^*)\,p(\mathbf{P}^*)}{p(\mathbf{Y} \mid \mathbf{P}^{(t)})\,p(\mathbf{P}^{(t)})} \frac{p(\mathbf{P}^{(t)})}{p(\mathbf{P}^*)}\right] \tag{33.a}$$

Hence, the Metropolis-Hastings ratio is given by the ratio of the likelihoods, that is,

$$\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p(\mathbf{Y} \mid \mathbf{P}^*)}{p(\mathbf{Y} \mid \mathbf{P}^{(t)})}\right] \tag{33.b}$$

Such as for the random walk proposal, candidates moving to regions of higher probability (in this case, the likelihood) are always accepted. Candidates moving to regions of lower likelihoods can be accepted as well, but with probability $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$. Although the probability $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ given by equation (33.b) does not involve the prior distribution, the Markov chain still depends on the prior since it is used to generate the candidates in this case. This kind of proposal can be very effective for parameters with small prior variances. On the other hand, it may lead to very small acceptance rates if the prior has large variances. Also, it cannot be applied to an improper prior with unlimited variance.

A Metropolis-Hastings algorithm with an adaptive proposal distribution was presented by Haario et al [34]. This algorithm is not Markovian, but results in ergodic distributions. In this adaptive algorithm, a Gaussian proposal with center at the sample of the current state, $\mathbf{P}^{(t)}$, is given by [34,35]:

$$q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \begin{cases} \mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{0.1^2}{N}\mathbf{I}\right) & t \le 2N \\ (1-\beta)\mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{2.38^2}{N}\boldsymbol{\Sigma}_t\right) + \beta\mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{0.1^2}{N}\mathbf{I}\right) & t > 2N \end{cases} \tag{34}$$

where $\mathrm{N}(\mathbf{a},\mathbf{B})$ is a Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{B}$, N is the number of parameters, $\mathbf{I}$ is the identity matrix and $\boldsymbol{\Sigma}_t$ is the covariance matrix of the posterior distribution up to the state t. The positive constant β ($0 < \beta < 1$) is used to promote the mixing between $\mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{2.38^2}{N}\boldsymbol{\Sigma}_t\right)$ and $\mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{0.1^2}{N}\mathbf{I}\right)$, in order to avoid that the algorithm halts if $\boldsymbol{\Sigma}_t$ is not well defined.

5.4. Metropolis-Hastings Algorithm with Sequential Sampling by Blocks of Parameters

Different modified versions of the Metropolis-Hastings algorithm can be found in the literature (see, for example, [29]). In particular, a modified version of the Metropolis-Hastings algorithm has been proposed for cases that involve groups of linearly dependent parameters [28,35]. In this modified version, the sampling procedure and the acceptance/rejection test are sequentially performed separately for each block of parameters, within one loop of the Metropolis-Hastings algorithm [28,35].

As an example, consider a case where the vector of parameters $\mathbf{P}$ is split into two blocks of parameters $\mathbf{P}_1$ and $\mathbf{P}_2$, that is, $\mathbf{P}^T = [\mathbf{P}_1^T \; \mathbf{P}_2^T]$. The Metropolis-Hastings algorithm with sequential sampling by blocks of parameters can then be summarized by the following steps:

1. Let $t = 0$ and start the Markov chains with the sample $\mathbf{P}^{(0)}$.
2. Sample candidates $\mathbf{P}_1^*$ from the proposal distribution $q_1(\mathbf{P}_1^* \mid \mathbf{P}_1^{(t)})$ for the vector $\mathbf{P}_1$ and make $\mathbf{P}_2^* = \mathbf{P}_2^{(t)}$.
3. Compute the Metropolis-Hastings ratio

$$\alpha_1(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p(\mathbf{P}^* \mid \mathbf{Y})\, q_1(\mathbf{P}_1^{(t)} \mid \mathbf{P}_1^*)}{p(\mathbf{P}^{(t)} \mid \mathbf{Y})\, q_1(\mathbf{P}_1^* \mid \mathbf{P}_1^{(t)})}\right] \tag{35.a}$$

4. Generate a random number with a uniform distribution in (0,1), $U_1 \sim \mathrm{U}(0,1)$.
5. If $U_1 \leq \alpha_1(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, make $\mathbf{P}_1^{(t+1)} = \mathbf{P}_1^*$. Otherwise, make $\mathbf{P}_1^{(t+1)} = \mathbf{P}_1^{(t)}$.
6. Sample candidates $\mathbf{P}_2^*$ from the proposal distribution $q_2(\mathbf{P}_2^* \mid \mathbf{P}_2^{(t)})$ for the vector $\mathbf{P}_2$ and make $\mathbf{P}_1^* = \mathbf{P}_1^{(t+1)}$.
7. Compute the Metropolis-Hastings ratio

$$\alpha_2(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p(\mathbf{P}^* \mid \mathbf{Y})\, q_2(\mathbf{P}_2^{(t)} \mid \mathbf{P}_2^*)}{p(\mathbf{P}^{(t)} \mid \mathbf{Y})\, q_2(\mathbf{P}_2^* \mid \mathbf{P}_2^{(t)})}\right] \tag{35.b}$$

8. Generate a random number with a uniform distribution in (0,1), $U_2 \sim \mathrm{U}(0,1)$.
9. If $U_2 \leq \alpha_2(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, make $\mathbf{P}_2^{(t+1)} = \mathbf{P}_2^*$. Otherwise, make $\mathbf{P}_2^{(t+1)} = \mathbf{P}_2^{(t)}$.
10. Let $t = t+1$ and return to step 2 in order to generate the sequence $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, ..., \mathbf{P}^{(n)}\}$.

6. Practical Issues regarding Markov Chain Monte Carlo (MCMC) Methods

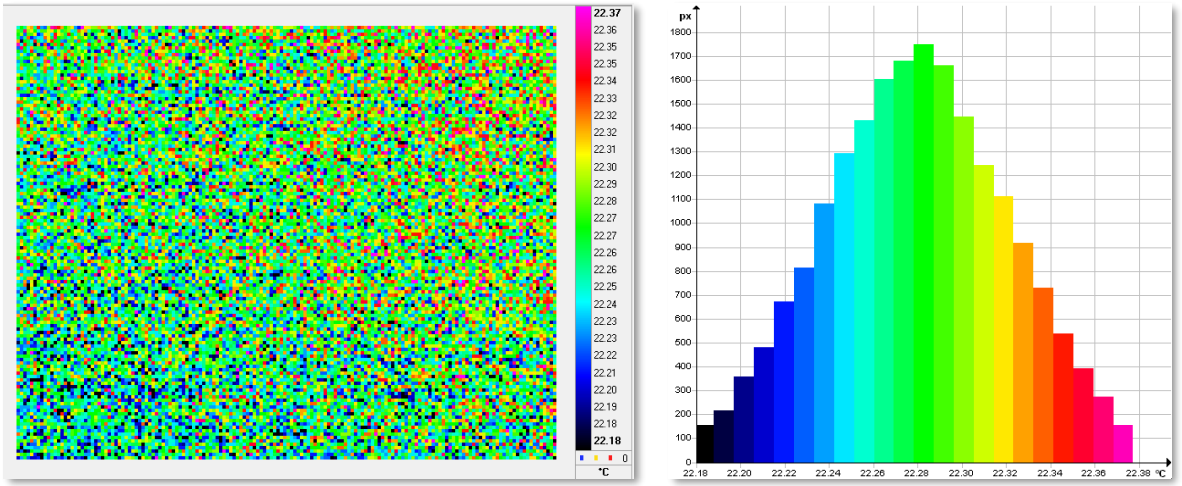
The objective of this section is to bring to the reader's attention some important aspects for the implementation of Markov Chain Monte Carlo methods. Although the discussion about likelihood and prior distributions is not limited to MCMC methods and is pertinent to Bayesian techniques in general, it was delayed until this section for the sake of organization of the text. Such is also the case regarding hierarchical models. In addition to these concepts, this section is also devoted to the analysis of the outputs of Markov chains.

6.1. Likelihood and Priors

The posterior distribution is proportional to the product of the likelihood function and the prior distribution (equation 17.b). As discussed in section 2, the likelihood function involves the solution of the mathematical formulation of the problem under analysis, that is, the solution of the direct or forward model, as well as the measurements and their related uncertainties. Measurement errors are modelled after the calibration of sensors and instruments used to collect the experimental data. The likelihood in section 2 was considered as Gaussian and given by equation (6.b). Such a model is in general appropriate for temperature measurements taken with thermocouples or infrared cameras. For example, figure 2.b presents the histogram of the readings (see figure 2.a) of a plate maintained at the constant temperature of 23 ºC, obtained with a SC7600 Flir infrared camera [36]. This histogram clearly approximates a Gaussian distribution. For other likelihood models more appropriate to different phenomena the reader is referred to [5].

A Gaussian prior was also considered in section 4, given by equation (18) for a multivariate case, with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$, denoted as $\mathbf{P} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$. For one single parameter P_j, a *Gaussian prior* with mean μ_j and variance σ_j^2, $P_j \sim \mathrm{N}(\mu_j, \sigma_j^2)$, is given by

$$p(P_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2} \frac{(P_j - \mu_j)^2}{\sigma_j^2} \right] \qquad \text{in} \quad -\infty < P_j < \infty \tag{36}$$



(a) (b)

Figure 2. (a) Thermal image with an infrared camera of an isothermal plate;
(b) Histogram of the temperature measurements [36].

Random variables modelled by the Gaussian prior have support in □. Hence, they may assume negative values, although this might happen with small probabilities depending on the values of μ_j and σ_j^2. On the other hand, several physical parameters only allow positive values, such as, for example, thermal conductivity, specific heat and thermal diffusivity.

A very simple prior that allows lower and upper bounds for the parameter values is the *Uniform distribution* $P_j \sim \mathrm{U}(a, b)$ given by

$$p(P_j) = \begin{cases} \dfrac{1}{(b-a)} & , \quad a < P_j < b \\ 0 & , \quad \text{elsewhere} \end{cases} \tag{37}$$

Mean and variance of the uniform distribution are given by $\frac{1}{2}(a+b)$ and $\frac{1}{12}(b-a)^2$, respectively.

In the uniform distribution, any value in $a < P_j < b$ is equally probable. If in this interval values around a known mean are more likely to occur than elsewhere, like in a Gaussian distribution, but the probability density is zero in $P_j \le a$ and $P_j \ge b$, one possible prior can be obtained by combining equations (36) in (37), which is called *truncated Gaussian distribution*, that is,

$$p(P_j) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\dfrac{1}{2} \dfrac{(P_j - \mu_j)^2}{\sigma_j^2} \right] & , \quad a < P_j < b \\ 0 & , \quad \text{elsewhere} \end{cases} \tag{38}$$

where $a < \mu_j < b$.

Other distributions that satisfy positive constraints are available. For example, the *Rayleigh distribution* $P_j \sim \mathrm{R}(\gamma_0)$ is given by:

$$p(P_j) = \frac{P_j}{\gamma_0^2} \exp\left[-\frac{1}{2} \left(\frac{P_j}{\gamma_0} \right)^2 \right] \qquad \text{for} \quad P_j > 0 \tag{39}$$

and depends only on the scale parameter (centerpoint) γ_0. The mean and the variance of Rayleigh's distribution are given by $\gamma_0 \sqrt{\dfrac{\pi}{2}}$ and $\dfrac{4-\pi}{2}\gamma_0^2$, respectively.

The *Gamma distribution* with parameters α and β, denoted as $P_j \sim \mathrm{G}(\alpha, \beta)$, has the following density:

$$p(P_j) = \frac{1}{\beta^\alpha \Gamma(\alpha)} P_j^{\alpha-1} \exp\left(-\frac{P_j}{\beta} \right) \quad \text{for} \quad P_j > 0 \tag{40}$$

with mean $\alpha\beta$ and variance $\alpha\beta^2$, where $\Gamma(\alpha)$ is the gamma function of argument α. For $\beta = 1$, the so-called one-parameter gamma distribution is obtained. The density that results by making $\alpha = 1$ is called exponential distribution.

The *Beta distribution* $P_j \sim \mathrm{Be}(\alpha, \beta)$ has support in $0 < P_j < 1$. The density of this distribution is given by:

$$p(P_j) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} P_j^{\alpha-1} (1 - P_j)^{\beta-1} \quad \text{in } 0 < P_j < 1 \tag{41}$$

with mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Figure 3 illustrates the probability distributions $\mathrm{U}(0,1)$, $\mathrm{N}(0.5, 0.5^2)$, $\mathrm{R}(0.5)$, $\mathrm{G}(1.5,1.5)$ and $\mathrm{Be}(1.5,1.5)$. These distributions were normalized by their maximum values to allow the comparison among them.

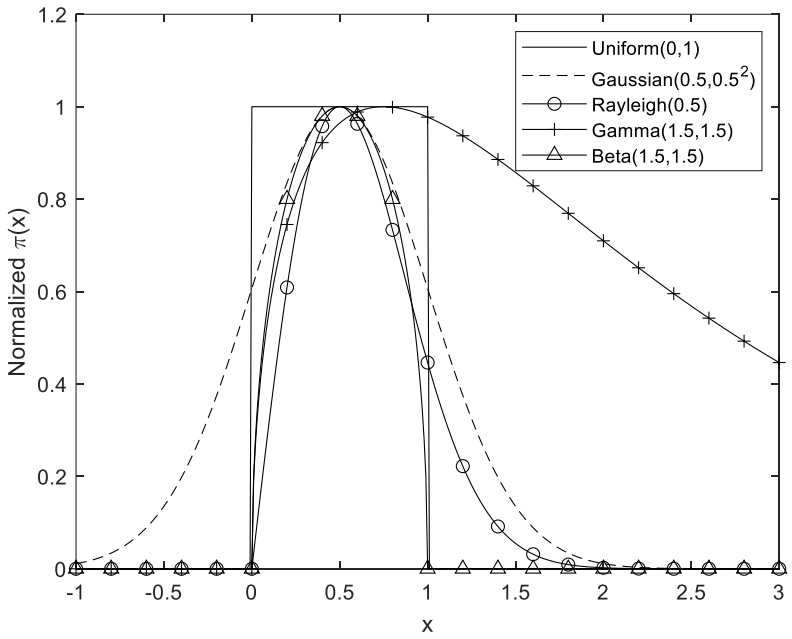


Figure 3. Probability distributions

The probability distributions given by equations (36) to (41) were written for one single random variable, but they can be easily extended for multivariate cases [1,4,5,8,28]. A multivariate prior is usually required for the solution of inverse problems in situations where the parameters represent point values of a function. Such is the case illustrated by Figure 1 for time varying functions. Another typical case involves spatially distributed functions, like a thermophysical property that varies within the medium.

The multivariate Gaussian distribution is given by equation (18). The use of Gaussian priors for function estimation is of great interest, because they tend to smooth out the oscillations in the solution caused by the ill-posed character of the inverse problem. If the parameters are independent, the prior covariance matrix is diagonal, with elements given by the variances, σ_i^2, that is,

$$\left[\mathbf{V}\right]_{i,j} = \begin{cases} \sigma_i^2 \ , \ i = j \\ 0 \ , \ i \neq j \end{cases} \tag{42}$$

However, rarely there is such independence in practice when the parameters are local function values. Consider, for example, a function that varies spatially. In this case, the parameters correspond to the mean values of the function inside finite volumes used for the discretization

of the spatial domain. The correlation between the parameters of different finite volumes must be taken into account in the covariance matrix of the prior information.

Works related to imaging have demonstrated that the *Matérn class* [21] of covariance functions may be appropriate for taking into account the correlation between spatially distributed parameters. The elements of the Matérn covariance matrix for the Gaussian prior distribution can be written as [21]:

$$[\mathbf{V}]_{i,j} = \begin{cases} \sigma_i^2 \quad , \quad i = j \\ \sigma_i^2 \dfrac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\dfrac{\sqrt{2\alpha}\left|\mathbf{r}_i - \mathbf{r}_j\right|}{l} \right)^{\alpha} K_\alpha \left(\dfrac{\sqrt{2\alpha}\left|\mathbf{r}_i - \mathbf{r}_j\right|}{l} \right) \quad , \quad i \neq j \end{cases} \tag{43}$$

where $\mathbf{r}_i$ is the position vector of the finite volume i, $\left|\mathbf{r}_i - \mathbf{r}_j\right|$ is the distance between finite volumes i and j, $\alpha > 0$ is a parameter that controls the smoothness of the random field, l is the characteristic length scale that controls the spatial range of correlation, Γ is the gamma function and K_α is the modified Bessel function of the second kind of order α.

With the Matérn covariance matrix, the correlation is more significant for neighbouring finite volumes and decreases for increasing distances between them. The correlation decay rate is controlled by the characteristic length scale l and the smoothness parameter α.

Markov Random Fields are also popular for priors in inverse problems of estimating spatially distributed functions or time varying functions [5]. A set $\{P_1, P_2, \ldots, P_N\}$ is a Markov Random Field if the conditional distribution of P_j depends only on the set of its neighbours [28].

A common use of a Markov Random Field is for priors that resemble Tikhonov's regularization [5], written in the following general form:

$$p(\mathbf{P}) \propto \exp\left[-\frac{1}{2}\gamma \left\| \mathbf{D}(\mathbf{P} - \tilde{\mathbf{P}}) \right\|^2 \right] \tag{44}$$

where ||.|| denotes the L_2 norm. The constant γ is a parameter associated with uncertainties in the prior and $\tilde{\mathbf{P}}$ is a reference value for $\mathbf{P}$. The vector $\tilde{\mathbf{P}}$ is commonly taken as zero without loss of generality. As for the matrix $\mathbf{D}$, it should be such that $\mathbf{D}(\mathbf{P} - \tilde{\mathbf{P}})$ involves the parameter P_j and its neighbors, in order to characterize a Markov random field. For cases that $\mathbf{P}$ represent local values of a one-dimensional function (such as a function varying in time or in one single spatial coordinate), the following matrices can be used [5]:

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \quad \text{with size } (N-1)\,\text{x}\,N \tag{45.a}$$

or

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \quad \text{with size } (N-2)\,\text{x}\,N \tag{45.b}$$

which are analogous to the matrices used in first-order and second-order Tikhonov´s regularization, respectively.

Equation (44) can be rewritten as:

$$p(\mathbf{P}) \propto \exp\left[-\frac{1}{2}\gamma(\mathbf{P}-\tilde{\mathbf{P}})^T \mathbf{Z}(\mathbf{P}-\tilde{\mathbf{P}}) \right] \tag{46}$$

where

$$\mathbf{Z} = \mathbf{D}^T\mathbf{D} \tag{47}$$

Equation (46) is in a form similar to that of a Gaussian distribution. For this reason, it is also called a Gaussian Markov Random Field [28] or a Gaussian Smoothness Prior [5]. By comparing equation (46) with the canonical Gaussian multivariate distribution, one can notice that the mean and the covariance matrix of this prior are given by $\tilde{\mathbf{P}}$ and $\gamma^{-1}\mathbf{Z}^{-1}$, respectively. Therefore, we can write the Gaussian Smoothness Prior as:

$$p(\mathbf{P}) = (2\pi)^{-N/2}\gamma^{N/2}\left|\mathbf{Z}^{-1}\right|^{-1/2} \exp\left[-\frac{1}{2}\gamma(\mathbf{P}-\tilde{\mathbf{P}})^T \mathbf{Z}(\mathbf{P}-\tilde{\mathbf{P}}) \right] \tag{48}$$

An important remark about this prior is that, with $\mathbf{D}$ given by equations (45.a,b), its variance is unbounded, since the matrix $\mathbf{Z}$ is singular and $\mathbf{Z}^{-1}$ does not exist. Densities with unbounded variances are denoted as *improper* [5,28].

We now discuss another Markov Random Field prior, which gives high probabilities for piecewise regular solutions with sparse gradients. The *Total Variation (TV) prior* satisfies these characteristics, being quite appropriate for spatially varying functions that contain large variations at few boundaries within the domain and with small variations within the regions limited by such boundaries [5]. The TV prior is given by [5]:

$$p(\mathbf{P}) \propto \exp\left[-\gamma\, TV(\mathbf{P})\right] \tag{49}$$

where

$$TV(\mathbf{P}) = \sum_{j=1}^{N} V_j(\mathbf{P}) \qquad ; \qquad V_j(\mathbf{P}) = \frac{1}{2}\sum_{i \in N_j} l_{ij} \left| P_i - P_j \right| \tag{50.a,b}$$

and N_j is the set of neighbors to P_j, while l_{ij} is the length of the edge between neighbors.

The TV prior is improper, such as the Gaussian smoothness prior. The representation of equation (49) in terms of a canonical probability density would require the derivation of an expression for the normalizing constant $\int_{\mathbb{R}^N} p(\mathbf{P})\, d\mathbf{P}$, or, at least, practical means for its computation.

Although improper priors need to be used with caution, they do not pose difficulties for the application of the Metropolis-Hastings algorithm, since the normalizing constants of such densities are cancelled when $\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$ is computed with equation (30). On the other hand, both the Gaussian smoothness prior and the TV prior involve an additional parameter γ that needs to be specified for the application of MCMC methods. The specification of a value for such parameter can be made by numerical experiments, by using simulated experimental data that serve as a reference for the inverse problem under analysis. On the other hand, if a parameter is not known it shall be regarded as part of the inference problem within the Bayesian framework of statistics, leading in the case of γ to the use of hierarchical (*hyperprior*) models, as described below.

6.2. Hierarchical Models

The parameter γ appearing in the Gaussian smoothness prior given by equation (46) can be treated as a *hyperparameter*, that is, be estimated as part of the inference problem [5]. Consider, for example, the *hyperprior density* for γ in the form of a Rayleigh distribution (see equation 39), where the scale parameter γ_0 needs to be chosen in advance. Therefore, the posterior distribution, with the Gaussian likelihood given by equation (6.b), can be written as:

$$p(\gamma, \mathbf{P} \,|\, \mathbf{Y}) \propto \gamma^{(N+2)/2} \exp\left\{ -\frac{1}{2}[\mathbf{Y} - \mathbf{T}(\mathbf{P})]^T \mathbf{W}^{-1} [\mathbf{Y} - \mathbf{T}(\mathbf{P})] - \frac{1}{2}\gamma (\mathbf{P} - \tilde{\mathbf{P}})^T \mathbf{Z} (\mathbf{P} - \tilde{\mathbf{P}}) - \frac{1}{2}\left(\frac{\gamma}{\gamma_0} \right)^2 \right\} \tag{51}$$

On the other hand, the parameter γ appearing in the TV prior given by equation (49) cannot be treated as a hyperparameter. Such is the case because the normalizing constant of such prior is of difficult calculation and also depends on γ. Therefore, without the computation of the normalizing constant for this case, the effects of γ as a hyperparameter would not be correctly accounted for in the posterior distribution.

6.3. Output Analysis

We basically follow references [22,28] for the material presented in this section and consider an analysis involving one single component P_j of the vector of parameters $\mathbf{P}$.

Let $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ be the set of samples of a homogeneous and reversible Markov chain with n states for the posterior distribution of P_j. The Markov chain should already have reached equilibrium before the samples can be tentatively used to represent the posterior distribution. The number of states required for the chain to reach equilibrium is denoted as the burn-in period. We consider that the burn-in period contains the first z states of the Markov chain, that is, the set of samples to be used to represent the posterior is $\{P_j^{(z+1)}, P_j^{(z+2)}, \ldots, P_j^{(n)}\}$ instead of $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$. For simplicity in notation, the index of samples after the burn-in period is changed from $t = z+1, \ldots, n$ to $r = 1, \ldots, s$, where $s = n - z$.

A function $f(P_j^{(s)})$ of the samples $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\}$ is called a *statistic* if it does not depend on any other unknown parameters. Some useful statistics are:

Minimum Value: $$f(P_j^{(s)}) = P_{j,\min}^{(s)} = \min\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\} \tag{52.a}$$

Maximum Value: $$f(P_j^{(s)}) = P_{j,\max}^{(s)} = \max\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\} \tag{52.b}$$

Median: $$f(P_j^{(s)}) = \tilde{P}_j^{(s)} = \text{med}\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\} \tag{52.c}$$

Mean: $$f(P_j^{(s)}) = \bar{P}_j^{(s)} = \frac{1}{s}\sum_{r=1}^{s} P_j^{(r)} \tag{52.d}$$

Variance: $$f(P_j^{(s)}) = \text{var}(P_j^{(s)}) = \frac{1}{s-1}\sum_{r=1}^{s}\left(P_j^{(r)} - \bar{P}_j^{(s)}\right)^2 \tag{52.e}$$

Since $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\}$ are realizations of a random variable, a statistic is itself a random variable as well. A statistic calculated with the samples will be a good representation of a statistic of the population if the samples are a good representation of the population. This certainly depends on the size s and on the independence of the samples. For example, if the Markov chain is ergodic the mean $\bar{P}_j^{(s)}$ provides a strongly consistent estimate of the mean of the limiting distribution as $s \to \infty$, that is,

$$\bar{P}_j^{(s)} \to E\left[P_j\right] \tag{53}$$

This result is the law of large numbers for a Markov chain.

If $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$ are independent samples, then the *variance of the mean* $\bar{P}_j^{(s)}$ is:

$$\mathrm{var}[\bar{P}_j^{(s)}] = \frac{\mathrm{var}(P_j^{(s)})}{s} \tag{54}$$

where $\mathrm{var}(P_j^{(s)})$ is the variance of $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$ (equation 52.e). On the other hand, since the samples are in general correlated, equation (54) is rewritten as:

$$\mathrm{var}[\bar{P}_j^{(s)}] = \frac{\tau_j \, \mathrm{var}(P_j^{(s)})}{s} \tag{55}$$

where τ_j is the *integrated autocorrelation time* (IACT) for parameter P_j, which represents the number of correlated samples between independent samples in the set $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$. Therefore, the effective number of independent samples in $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$ is $s_{eff,j} = s / \tau_j$.

The *autocovariance function of lag k* of the chain for the parameter P_j is defined by:

$$C_j(k) = \mathrm{cov}[P_j^{(r)}, P_j^{(r+k)}] \tag{56}$$

Clearly, the variance of $P_j^{(r)}$ is $C_j(0)$.

The *normalized autocovariance function of lag k* is given by:

$$\rho_j(k) = \frac{C_j(k)}{C_j(0)} \tag{57}$$

so that $\rho_j(0) = 1$. Hence, $P_j^{(r)}$ is perfectly correlated with itself. The calculation of the normalized autocovariance function is straightforward, since several computational packages have functions available for such a purpose.

The integrated autocorrelation time is related to the normalized autocovariance function by:

$$\tau_j = 1 + 2\sum_{k=1}^{\infty} \rho_j(k) \tag{58}$$

For the calculation of τ_j, the summation in equation (58) needs to be truncated at a finite number of terms $s^* \leq s$. In fact, $\rho_j(k)$ is expected to tend to zero as k increases, as it is dominated by noise for large k. Therefore, s^* can be selected by increasing k until $\rho_j(k)$ approaches zero, thus avoiding the terms that are dominated by noise.

For s sufficiently large and for a uniformly ergodic Markov chain, the distribution of $\dfrac{\bar{P}_j^{(s)} - E\left[P_j\right]}{\sqrt{\mathrm{var}[\bar{P}_j^{(s)}]}}$ tends to a standard Gaussian distribution, with zero mean and unitary standard deviation. Thus,

$$\frac{\bar{P}_j^{(s)} - E\left[P_j\right]}{\sqrt{\mathrm{var}[\bar{P}_j^{(s)}]}} \overset{d}{\rightarrow} \mathrm{N}(0,1) \quad \text{as} \quad s \rightarrow \infty \tag{59}$$

where $\overset{d}{\rightarrow}$ indicates that the distribution of the random variable on the left tends to the distribution on the right and $\mathrm{var}[\bar{P}_j^{(s)}]$ is obtained with equation (55). Equation (59) is a statement of the central limit theorem of the distribution of $\bar{P}_j^{(s)}$. Therefore, the mean of the samples in the Markov chain can be reported with related uncertainties as $\bar{P}_j^{(s)} \pm \eta\sqrt{\mathrm{var}[\bar{P}_j^{(s)}]}$, where η is a constant that defines the approximate confidence interval of $\bar{P}_j^{(s)}$ (η = 2.576 for a 99% confidence interval if s is large).

The statistical efficiency of the sampling algorithm can be assessed by examining τ_j for each parameter P_j , j = 1,...,N. Algorithms that result in small values of τ_j promote better sampling. For cases involving many parameters, the statistical efficiency can be examined with the integrated autocorrelation time of the posterior distribution $\pi(\mathbf{P}^{(r)} \mid \mathbf{Y})$, $r = 1,...,s$ [35].

Quantitative techniques are available for the analysis of the convergence of a Markov chain to an equilibrium distribution. Geweke´s technique [37] compares the means calculated with the samples of different ranges of states of the Markov chain. Let:

$$\bar{P}_j^a = \frac{1}{s_a}\sum_{r=1}^{s_a} P_j^{(r)} \quad \text{and} \quad \bar{P}_j^b = \frac{1}{(s - s_b)}\sum_{r=s_b}^{s} P_j^{(r)} \tag{60.a,b}$$

be the means calculated with s_a and $(s - s_b)$ states, respectively. Geweke [37] recommended:

$$s_a = 0.1s \quad \text{and} \quad s_b = 0.5s + 1 \tag{61.a,b}$$

that is, the means of the samples of the first 10% and of the last 50% of the states in the Markov chain are compared. If an equilibrium distribution is reached, $\left|\bar{P}_j^a - \bar{P}_j^b\right| \approx 0$.

For the convergence analysis, it is also recommended to repeat the sampling procedure by starting the Markov chains from different initial values. Gelman and Rubin [38] developed a method for inference on multiple chains, based on two steps: (i) An estimate is obtained for

the posterior distribution with an initial Markov chain, which is then used to start new independent chains. The initial states for these new multiple chains must have a dispersion larger than that of the initial chain; (ii) The new multiple chains are then used for inference with analyses inter chains and within each chain. The posterior distribution simulated with the multiple chains exhibit a variability larger than that of the initial chain.

The multiple chains also allow a convergence analysis to verify if an equilibrium distribution has been reached to represent the sought posterior. We consider the case of a parameter P_j , $j = 1,...,N$.

The variance of the means of m chains, each one with s states, is given by [38]:

$$B_j = \frac{1}{(m-1)} \sum_{k=1}^{m} \left(\bar{P}_j^k - \bar{P}_j \right)^2 \tag{62}$$

where $\bar{P}_j^k$ is the mean of the chain k, $k = 1,...,m$, and $\bar{P}_j$ is the mean of these means.

The mean of the m variances of the chains $k = 1,...,m$, is given by [38]:

$$W_j = \frac{1}{m(s-1)} \sum_{k=1}^{m} \sum_{r=1}^{s} \left(P_j^{(r),k} - \bar{P}_j^k \right)^2 \tag{63}$$

where $P_j^{(r),k}$ is the sample for P_j at state r, $r = 1,\ldots,s$, of chain k, $k = 1,...,m$.

The variance of the posterior distribution simulated with the multiple chains for P_j is thus obtained as [38]:

$$\hat{\sigma}_j^2 = \left(1 - \frac{1}{s} \right) W_j + \frac{1}{s} B_j \tag{64}$$

The variance of the total number of samples of the multiple chains, $\hat{\sigma}_j^2$, overestimate the variance of the actual posterior while the equilibrium distribution has not been reached. On the other hand, W_j underestimates the variance of the actual posterior if each chain has not reached equilibrium. Gelman and Rubin [38] thus proposed a parameter to indicate convergence based on $\hat{\sigma}_j^2$ and W_j, called scale reduction coefficient, which was simplified by Gamerman and Lopes [28] and is given by:

$$\hat{R}_j = \sqrt{\frac{\hat{\sigma}_j^2}{W_j}} \tag{65}$$

Note that $\hat{R}_j > 1$, but $\hat{R}_j \to 1$ when $s \to \infty$. Gelman and Shirley [39] have suggested $\hat{R}_j < 1.1$ as the convergence test of the multiple chains, but larger threshold values have also been proposed [28].

7. Reduction of the Computational Time for Markov Chain Monte Carlo (MCMC) Methods

For many practical cases the direct problem solution with the high-fidelity model is very time consuming. Limitations are then imposed on the number of states of the Markov chains that can be computed within a feasible time, which can make the use of standard MCMC methods impractical, especially when the number of unknown parameters is large. One possible way to overcome such difficulties is to use a low-fidelity model, instead of the high-fidelity model, for the computation of the direct problem solution at each state of the Markov chain. However, low-fidelity models reproduce the observed data (measurements) with uncertainties larger than those of the high-fidelity model. Therefore, different approaches have been developed in order to improve the solution of inverse problems obtained with low-fidelity models, including the Delayed Acceptance Metropolis-Hastings (DAMH) algorithm [40] and the Approximation Error Model (AEM) [5,41-45].

In the DAMH algorithm [40], the Metropolis-Hastings (MH) algorithm is regularly applied with the low-fidelity model. If a proposal sample is accepted with the low-fidelity model, another test of Metropolis-Hastings is performed with the high-fidelity model to finally decide if such sample should be accepted or not. In this sense, the DAMH can be seen as two nested Metropolis-Hastings algorithms, where the outer loop acts as a filter to pre-evaluate proposal candidates with the low-fidelity model. In the AEM approach [5,41-45], the statistical model of the approximation error between the high-fidelity and the low-fidelity models is constructed based on the prior distribution, and then represented as additional noise in the likelihood function for the solution of the inverse problem. It should be noted that there is a fundamental difference between the DAMH and the AEM approaches. While the AEM uses the posterior modified by the approximation error, the DAMH algorithm generates samples from the correct posterior [46].

7.1. Delayed Acceptance Metropolis-Hastings (DAMH) Algorithm

The DAMH algorithm can be summarized as follows [40]:

1. Let $t = 0$ and start the Markov chain with the sample $\mathbf{P}^{(0)}$ at the initial state.
2. Sample a candidate point $\mathbf{P}^*$ from a proposal distribution $q(\mathbf{P}^* \mid \mathbf{P}^{(t)})$.
3. Calculate the probability $\alpha_{app}(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ by using the low-fidelity model, where

$$\alpha_{app}(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p_{app}(\mathbf{P}^* \mid \mathbf{Y})\, q(\mathbf{P}^{(t-1)} \mid \mathbf{P}^*)}{p_{app}(\mathbf{P}^{(t-1)} \mid \mathbf{Y})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t-1)})}\right] \tag{66.a}$$

4. Generate a random value $U_{app} \sim \mathrm{U}(0,1)$.
5. If $U_{app} \le \alpha_{app}(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, proceed to step 6. Otherwise, return to step 2.
6. Calculate a new acceptance factor with the high-fidelity model:

$$\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{p(\mathbf{P}^* \mid \mathbf{Y})\, q(\mathbf{P}^{(t-1)} \mid \mathbf{P}^*)}{p(\mathbf{P}^{(t-1)} \mid \mathbf{Y})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t-1)})}\right] \tag{66.b}$$

7. Generate a new random value $U \sim \mathrm{U}(0,1)$.
8. If $U \le \alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ set $\mathbf{P}^{(t+1)} = \mathbf{P}^*$. Otherwise, set $\mathbf{P}^{(t+1)} = \mathbf{P}^{(t)}$.
9. Make $t = t+1$ and return to step 2 in order to generate the sequence $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(n)}\}$.

where $p_{app}(\mathbf{P} \mid \mathbf{Y})$ and $p(\mathbf{P} \mid \mathbf{Y})$ are the posterior distributions with likelihoods computed with the low-fidelity model and with the high-fidelity model, respectively.

The DAMH algorithm is expected to take advantage of the fast computations of the low-fidelity model in order to find, in step 5, possible candidates to be accepted with the high-fidelity model in step 8. The DAMH algorithm can be quite effective, especially in the case of a low acceptance ratio of the Metropolis-Hastings algorithm. Therefore, depending on how fast the solution of the low-fidelity model is as compared to that of the high-fidelity model, as well as on the acceptance ratio, the use of the DAMH algorithm might result in significant reductions in computational times as compared to those from the regular Metropolis-Hastings algorithm applied with the high-fidelity model.

7.2. Approximation Error Model (AEM) Approach

In the Approximation Error Model (AEM) approach, the statistical model of the approximation error is constructed and then represented as additional noise in the measurement model [5,41-45]. With the hypotheses that the measurement errors are additive and independent of the parameters $\mathbf{P}$ we can write:

$$\mathbf{Y} = \mathbf{T}(\mathbf{P}) + \boldsymbol{\varepsilon} \tag{67}$$

where $\mathbf{T}(\mathbf{P})$ is a quite accurate solution of the high-fidelity direct (forward) model. The vector of measurement errors, $\boldsymbol{\varepsilon}$, are assumed here to be Gaussian, with zero mean and known covariance matrix $\mathbf{W}$, so that the likelihood function is given by equation (6.b).

Let $\mathbf{T}_{app}(\mathbf{P})$ be the solution of a low-fidelity model that is used for the solution of the inverse problem in place of the high-fidelity model, $\mathbf{T}(\mathbf{P})$. Equation (67) can be re-written as:

$$\mathbf{Y} = \mathbf{T}_{app}(\mathbf{P}) + [\mathbf{T}(\mathbf{P}) - \mathbf{T}_{app}(\mathbf{P})] + \boldsymbol{\varepsilon} \tag{68}$$

By defining the approximation error between the high-fidelity and the low-fidelity model solutions as:

$$\mathbf{e}(\mathbf{P}) = [\mathbf{T}(\mathbf{P}) - \mathbf{T}_{app}(\mathbf{P})] \tag{69}$$

equation (68) can be written as:

$$\mathbf{Y} = \mathbf{T}_{app}(\mathbf{P}) + \boldsymbol{\eta}(\mathbf{P}) \tag{70}$$

where

$$\boldsymbol{\eta}(\mathbf{P}) = \mathbf{e}(\mathbf{P}) + \boldsymbol{\varepsilon} \tag{71}$$

One difficult with such an approach is to model the total error $\boldsymbol{\eta}(\mathbf{P})$, which includes the direct problem approximation error, $\mathbf{e}(\mathbf{P})$, as well as the experimental error, $\boldsymbol{\varepsilon}$. A simple, but very effective approach is to model the approximation error as a Gaussian variable [5,41-45]. Another important point for the implementation of the approximation error model is that the statistics of $\boldsymbol{\eta}(\mathbf{P})$, like its mean and covariance matrix, are computed before the estimation procedure, based on the prior distribution of the model parameters [5,41-45]. Therefore, the use of the approximation error model with improper priors is not possible, since they exhibit unbounded variances.

Consider, for instance, a Gaussian prior and a Gaussian likelihood, given by equations (18) and (6.b), respectively. By using the approximation error model approach, the posterior distribution is given by [41]:

$$p(\mathbf{P}|\mathbf{Y}) \propto \exp\left\{-\frac{1}{2}[\mathbf{Y} - \mathbf{T}_{app}(\mathbf{P}) - \bar{\boldsymbol{\eta}}]^T \tilde{\mathbf{W}}^{-1}[\mathbf{Y} - \mathbf{T}_{app}(\mathbf{P}) - \bar{\boldsymbol{\eta}}] - \frac{1}{2}(\mathbf{P} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{P} - \boldsymbol{\mu})\right\} \tag{72}$$

where

$$\bar{\boldsymbol{\eta}} = \bar{\boldsymbol{\varepsilon}} + \bar{\mathbf{e}} + \boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}}\mathbf{V}^{-1}(\mathbf{P} - \boldsymbol{\mu}) \tag{73.a}$$

$$\tilde{\mathbf{W}} = \mathbf{W}_e + \mathbf{W} - \boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}}\mathbf{V}^{-1}\boldsymbol{\Gamma}_{\mathbf{P}\boldsymbol{\eta}} \tag{73.b}$$

and $\bar{\boldsymbol{\varepsilon}}$ and $\bar{\mathbf{e}}$ are the means of $\boldsymbol{\varepsilon}$ and **e,** respectively, while $\mathbf{W}_e$ is the covariance of **e** and $\boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}}$ is the covariance of $\boldsymbol{\eta}$ and $\mathbf{P}$. Equations (73.a,b) give the *complete error model* [41]. We note that, with the standard hypotheses regarding the measurement errors made above, $\bar{\boldsymbol{\varepsilon}} = 0$. By further neglecting the dependency of $\boldsymbol{\eta}$ and $\mathbf{P}$, that is, $\boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}} = \boldsymbol{\Gamma}_{\mathbf{P}\boldsymbol{\eta}} = 0$, equations (73.a,b) simplify to the so-called *enhanced error model*:

$$\bar{\boldsymbol{\eta}} \approx \bar{\mathbf{e}} \tag{74.a}$$

$$\tilde{\mathbf{W}} \approx \mathbf{W}_e + \mathbf{W} \tag{74.b}$$

Further details of the AEM approach are presented in Tutorial 14 of this METTI School.

References

[1] Lee, P. M., 2004, *Bayesian Statistics*, Oxford University Press, London.

[2] Bayes, T., 1763, An Essay towards Solving a Problem in the Doctrine of Chances, by the late Rv. Mr. Bayes, F.R.S. Communicated by Mr. Price in a Letter to John Cannon, A.M.R.F.S., *Phil. Trans*. 1763 53, 370-418, 1763, doi:10.1098/rstl.1763.0053

[3] Silver N., 2012, *The Signal and the Noise*, Penguin Press, New York.

[4] Winkler, R., 2003, *An Introduction to Bayesian Inference and Decision*, Probabilistic Publishing, Gainsville.

[5] Kaipio, J. and Somersalo, E., 2004, *Statistical and Computational Inverse Problems*, Applied Mathematical Sciences 160, Springer-Verlag.

[6] Beck, J. V., Blackwell, B. and St. Clair, C. R., 1985, *Inverse Heat Conduction: Ill-Posed Problems*, Wiley Interscience, New York.

[7] Tikhonov, A. N. and Arsenin, V. Y., 1977, *Solution of Ill-Posed Problems*, Winston & Sons, Washington, DC.

[8] Beck, J. V. and Arnold, K. J., 1977, *Parameter Estimation in Engineering and Science*, Wiley Interscience, New York .

[9] Alifanov, O. M., 1994, *Inverse Heat Transfer Problems*, Springer-Verlag, New York.

[10] Alifanov, O. M., Artyukhin, E. and Rumyantsev, A., 1995, *Extreme Methods for Solving Ill-Posed Problems with Applications to Inverse Heat Transfer Problems*, Begell House, New York.

[11] Woodbury, K., 2002, *Inverse Engineering Handbook*, CRC Press, Boca Raton.

[12] Sabatier, P. C., 1978, *Applied Inverse Problems*, Springer Verlag, Hamburg.

[13] Morozov, V. A., 1984, *Methods for Solving Incorrectly Posed Problems*, Springer Verlag, New York.

[14] Murio, D. A., 1993, *The Mollification Method and the Numerical Solution of Ill-Posed Problems*, Wiley Interscience, New York.

[15] Trujillo, D. M. and Busby, H. R., 1997, *Practical Inverse Analysis in Engineering*, CRC Press, Boca Raton.

[16] Hensel, E., 1991, *Inverse Theory and Applications for Engineers*, Prentice Hall, New Jersey.

[17] Kurpisz, K. and Nowak, A. J., 1995, *Inverse Thermal Problems*, WIT Press, Southampton, UK.

[18] Vogel, C., 2002, *Computational Methods for Inverse Problems*, SIAM, New York.

[19] Yagola A. G., Kochikov, I.V., Kuramshina, G. M. and Pentin, Y. A., 1999, *Inverse Problems of Vibrational Spectroscopy,* VSP, Netherlands.

[20] Calvetti, D., Somersalo, E., 2007, *Introduction to Bayesian Scientific Computing*, Springer, New York.

[21] Ozisik, M.N. and Orlande, H.R.B., 2021, *Inverse Heat Transfer: Fundamentals and Applications – 2nd Edition*, Boca Raton, CRC Press.

[22] Tan, S., Fox, C., and Nicholls, G., 2006, *Inverse Problems, Course Notes for Physics 707*, University of Auckland.

[23] A. Tarantola, 1987, *Inverse Problem Theory*, Elsevier.

[24] M. Bertero, P. Boccacci, 1998, *Introduction to Inverse Problems in Imaging*, Institute of Physics.

[25] Orlande, H., Fudym, F., Maillet, D., Cotta, R., 2011, *Thermal Measurements and Inverse Techniques*, CRC Press, Boca Raton.

[26] Jari P. Kaipio & Colin Fox, 2011, The Bayesian Framework for Inverse Problems in Heat Transfer, *Heat Transfer Engineering*, 32:9, 718-753.

[27] Orlande, H. R. B., 2012, Inverse Problems in Heat Transfer: New Trends on Solution Methodologies and Applications, *Journal of Heat Transfer*, v.134, p.031011.

[28] Gamerman, D. and Lopes, H.F., 2006, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall/CRC, 2nd edition, Boca Raton.

[29] Brooks, S., Gelman, A., Jones, G., Meng, X, 2011, *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, Boca Raton.

[30] McGrayne, S. B., 2011, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, Devon.
[31] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953, Equation of State Calculation by Fast Computing Machines, *J. Chemical Phys.*, vol. 21, pp. 1087-1092
[32] Hastings, W. K., 1970, Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, vol. 57, pp. 97-109.
[33] http://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm, consulted on July 07, 2023.
[34] Haario, H., Saksman, E., Tamminen, J., 2001, An Adaptive Metropolis Algorithm, *Bernoulli*, vol. 7, pp. 223-242.
[35] Cui, T., 2010, *Bayesian Calibration of Geothermal Reservoir Models via Markov Chain Monte Carlo*, Ph.D. Thesis, The University of Auckland.
[36] Fonseca, H.M., Orlande, H.R.B., Fudym, O., Sepúlveda, F., 2014, A statistical inversion approach for local thermal diffusivity and heat flux simultaneous estimation, *Quantitative InfraRed Thermography*, pp. 170-189.
[37] Geweke, J., 1992, Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, in *Bayesian Statistics*, Bernardo, J., Berger, J., Dawid, a., Smith, A., (eds)., Oxford University Press
[38] Gelman, A., Rubin, D., 1992, Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, vol. 7, pp. 457-472.
[39] Gelman, A., Shirley, K., 2011, *Inference from Simulations and Monitoring Convergence*, Chapter 6 in Brooks, S., Gelman, A., Jones, G., Meng, X., 2011, Handbook of Markov Chain Monte Carlo, CRC Press, Boca Raton
[40] Christen, J., Fox, C., Markov chain Monte Carlo Using an Approximation, Journal of Computational and Graphical Statistics, vol. 14, no. 4, pp. 795–810, 2005.
[41] Nissinen, A., 2011, *Modelling Errors in Electrical Impedance Tomography*, Dissertation in Forestry and Natural Sciences, University of Eastern Finland.
[42] Nissinen, A., Heikkinen, L., Kaipio. J., 2008, The Bayesian approximation error approach for electrical impedance tomography – experimental results, *Meas. Sci. Technology*, vol. 19., pp. 015501.
[43] Nissinen, A., Heikkinen, L.. Kolehmainen, V., Kaipio. J., 2009, Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography, *Meas. Sci. Technology*, vol. 20, pp. 105504.
[44] Nissinen, A., Kolehmainen, V., Kaipio. J., 2011, Compensation of modeling errors due to unknown boundary domain in electrical impedance tomography, *IEEE Trans. Med. Im*., vol. 30, pp. 231-242.
[45] Nissinen, A., Kolehmainen, V., Kaipio. J., 2011, Reconstruction of domain boundary and conductivity in electrical impedance tomography using the approximation error approach, *Int. J. Uncertainty Quant.*, vol. 1, pp. 203–222.
[46] Orlande, H. R. B., Dulikravich, G. S., Neumayer, M., Watzenig, D., Colaço, M., 2014, Accelerated Bayesian Inference for the Estimation of Spatially Varying Heat Flux in a Heat Conduction Problem, *Numerical Heat Transfer, Part A: Applications*, vol. 65, pp. 1-25

Invited Conference.

Contactless thermal measurements using MRI : applications in interventional radiology and perspectives in pathophysiology.

Valery Ozenne[1 2 3]
[1] CNRS, CRMSB, UMR 5536, IHU Liryc, Université de Bordeaux, Bordeaux, France.

E-mail: valery.ozenne@u-bordeaux.fr

Abstract. Interventional thermoablation procedures are used to destroy irreversibly pathological tissues (tumor cells, etc.) by means of localized energy deposition. The procedure, whatever the modality and medical device envisaged, is divided into three phases: i) ballistics, ii) energy deposition iii) post-ablation assessment. During ablation, it is rare to be able to objectify the energy delivered. Procedures are based on the manufacturer's abacus (power and emission time) for the devices used, and not on the temperature rise obtained locally in the targeted tissues. As a result, personalization of the treatement is impossible, and monitoring of the procedure's progress is limited. The unique properties of MRI make it capable to map volumetric temperature changes in real time during an ablation and predict the final lesion size. The advantages of this technology is relevant with regard to efficacy and safety of the procedure.

The talk will introduce Magnetic resonance thermometry (MRT), a non-invasive technique for monitoring volumetric tissue temperature in real-time. Several applications will be presented at different stages of clinical advancement (from bench to clinical trial). Finally, the adaptation of these methods to study non-invasively the thermoregulatory mechanisms in the human body will be presented, along with links with the SFT community.

https://metti8.sciencesconf.org/

The 8th edition of the Advanced Autumn school 'Thermal Measurement and Inverse Techniques' is run by the METTI Group (**ME**sures en **T**hermique et **T**echniques **I**nverses) that constitutes a division of the Société Française de Thermique (SFT, French Heat Transfer Society).

* * *

Finding 'causes' from measured 'consequences' using a mathematical model linking the two is an inverse problem. This is met in different areas of physical sciences, especially in Heat Transfer. Techniques for solving inverse problems as well as their applications may seem quite obscure for newcomers to the field. Experimentalists desiring to go beyond traditional data processing techniques for estimating the parameters of a model with the maximum accuracy feel often ill prepared in front of inverse techniques. In order to avoid biases at different levels of this kind of involved task, it seems compulsory that specialists of measurement inversion techniques, modelling techniques and experimental techniques share a wide common culture and language. These exchanges are necessary to take into account the difficulties associated to all these fields. It is in this state of mind that this school is proposed. The METTI Group (Thermal Measurements and Inverse Techniques), which is a division of the French Heat Transfer Society (SFT), has already run or co- organized seven similar schools, in the Alps (Aussois, 1995 and 2005), in the Pyrenees (Bolquère-Odeillo, 1999), in Brasil (Rio de Janeiro, 2009), in Bretagne (Roscoff, 2011), in Pays Basque (Biarritz, 2015) and in Porquerolles island (Porquerolles 2019). For this eighth edition the school is again open to participants from the European Community with the support of the Eurotherm Committee.

* * *

Two books are distributed at the beginning of the school. Volume 1 contains the texts used as supports for the lectures and Volume 2 contains the texts used as supports for the tutorials.