DSP

DIGITAL SIGNAL AND IMAGE PROCESSING SERIES

Bayesian Approach to Inverse Problems

Edited by Jérôme Idier

ISTE

WILEY

Chapter 1

Inverse Problems, Ill-posed Problems

1.1. Introduction

In many fields of applied physics, such as optics, radar, heat, spectroscopy, geophysics, acoustics, radioastronomy, non-destructive evaluation, biomedical engineering, instrumentation and imaging in general, we are faced with the problem of determining the spatial distribution of a scalar or vector quantity – we often talk about an *object* – from direct measurements – called an *image* – or indirect measurements – called *projections* in the case of tomography, for example – of this object. Solving such imaging problems can habitually be broken down into three stages [HER 87, KAK 88]:

– a *direct problem* where, knowing the object and the observation mechanism, we establish a mathematical description of the data observed. This model needs to be accurate enough to provide a correct description of the physical observation phenomenon and yet simple enough to lend itself to subsequent digital processing;

– an *instrumentation problem* in which the most informative data possible must be acquired so that the imaging problem can be solved in the best conditions;

– an *inverse problem* where the object has to be estimated from the preceding model and data.

Obtaining a good estimate of the object obviously requires these three sub-problems to be studied in a coordinated way. However, the characteristic that these image reconstruction or restoration problems have in common is that they are often ill-posed or ill-conditioned. Higher level problems that are found in computer vision, such as

---

Chapter written by Guy DEMOMENT and Jérôme IDIER.

image segmentation, optical flow processing and shape reconstruction from shading, are also inverse problems and suffer from the same difficulties [AND 77, BER 88, MAR 87]. In the same way, a problem such as spectral analysis, which has similarities with the Fourier synthesis used in radio-astronomy, for example, and which is not usually treated as an inverse problem, can gain from being approached this way, as we will see later.

Schematically, there are two broad communities that are interested in these inverse problems from a methodological point of view:

– the *mathematical physics* community, with the seminal works of Phillips, Twomey and Tikhonov in the 1960s [PHI 62, TIK 63, TWO 62]. Sabatier was one of the pioneers in France [SAB 78]. A representative journal is *Inverse Problems*;

– the *statistical data processing* community, which can be linked to the work of Franklin in the late 1960s [FRA 70], although the ideas involved – the basis of Wiener filtering – had been bubbling beneath the surface in many works for several years [FOS 61]. The Geman brothers gave a major boost to image processing about twenty years ago [GEM 84] A representative journal is *IEEE Transactions on Image Processing*.

A very rough distinction can be made between these two communities by saying that the former deals with the problem in an infinite dimension, with the questions of existence, uniqueness and stability, which become very complicated for nonlinear direct problems, and solves it numerically in finite dimensions, while the latter starts with a problem for which the discretization has already been performed and is not called into question, and takes advantage of the finite nature of the problem to introduce prior information built up from probabilistic models.

In this chapter, we propose to use a basic example to point out the difficulties that arise when we try to solve these inverse problems.

1.2. Basic example

We will now illustrate the basic concepts introduced in this chapter by an artificial example that mixes the essential characteristics of several types of inverse problems.

We are looking for a *spectrum*, the square of the modulus of a function $\widehat{x}(\nu)$, $\nu \in \mathbb{R}$ but, because of the experimental constraints, we only have access to the dual domain of the variable ν, through the function $x(t)$ of which $\widehat{x}(\nu)$ is the Fourier transform (FT). What is more, imperfections in the apparatus mean that the function $x(t)$ is only observable as weighted by a "window" $h(t)$, which gives the observable function $y(t)$:

$$y(t) = h(t)\, x(t)\,. \tag{1.1}$$

To make our ideas clear, let us think of a visible optical interferometry device like that by Michelson. To have access to the emission spectrum of the light source, we measure an energy flux as a function of the phase difference between two optical paths. The interferogram obtained is, ignoring the additional constant, the Fourier transform of the function we are looking for but the limitations of the apparatus make the interferogram observable only in a limited area of space, which is equivalent to its being modulated by a weighting function $h(t)$. This is assumed to be known but the experimental data that is actually available is made up of a finite number of regularly-spaced samples of the function $y(t)$, which inevitably contain measuring errors that we assume to be additive. If we take a unit sampling step, we can write:

$$y_n = h_n\, x_n + b_n\,, \qquad n = 1, \ldots,\, N, \tag{1.2}$$

where y_n designates the available data, h_n and x_n the samples of the functions $h(t)$ and $x(t)$ respectively, and b_n the measurement "noise". This is a special case of a system of linear equations of the form:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{b} \tag{1.3}$$

that we will find repeatedly throughout this book. Here we have a diagonal matrix $\mathbf{A}$ which, at first glance, appears to be a simple situation.

A first difficulty appears, however, independently of the presence of the weighting $h(t)$: the discrete nature of the data means that we only have information on $\widehat{x}_1(\nu)$, $\nu \in [0, 1]$, a 1-periodic function deduced from $\widehat{x}(\nu)$ by the periodization due to the sampling, since we have:

$$x_n = \int_0^1 \widehat{x}_1(\nu)\, \exp\{2j\pi\nu n\}\; d\nu\,. \tag{1.4}$$

The samples x_n are in fact the Fourier series development coefficients of $\widehat{x}_1$. To have any hope of accessing $\widehat{x}$, it is necessary for $\widehat{x}(\nu)$ to have limited support and for the sampling step to be such that there is no aliasing. Observation model (1.2) can thus be written indifferently:

$$y_n = \int_0^1 \widehat{h} \star \widehat{x}_1(\nu)\, \exp\{2j\pi\nu n\}\; d\nu + b_n\,, \tag{1.5}$$

where $\widehat{h}(\nu)$ is the FT of $h(t)$. The presence of this convolution core expresses the loss of resolving power of the instrument due to the weighting by $h(t)$.

A simulated example is presented in Figure 1.1. Signal $x(t)$ is composed of three sine waves, the spectrum of which is marked by the circles in Figure 1.1a. Two have frequencies that are close together (relative frequency difference less than 0.008). Response $\widehat{h}(\nu)$ is a Gaussian of standard deviation $\sigma_{\widehat{h}} = 0.0094$ intentionally chosen

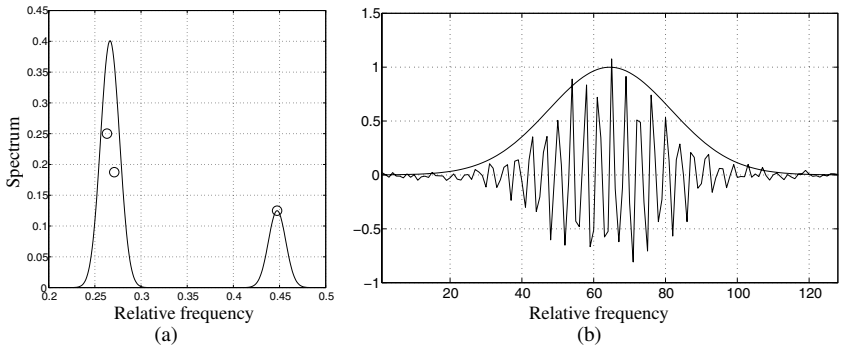

Figure 1.1. (a) *Spectrum $\widehat{x}$ of a linear combination of three sine waves, indicated by circles, and $\widehat{h} \star \widehat{x}$, where $\widehat{h}$ is a Gaussian of standard deviation close to* 0.01 *in relative frequency;* (b) *128 data points $\boldsymbol{y}$ simulating an interferogram that contains noise and is quantified, corresponding to model* (1.5)

high to point out clearly the difficulties of inversion. The "non-resolved" spectrum $\widehat{h} \star \widehat{x}$ is also represented in Figure 1.1a. Figure 1.1b superposes the $N = 128$ simulated data y_n and the series of weighting coefficients h_n, which also have a Gaussian form (of standard deviation $1/2\pi\sigma_{\widehat{h}} = 17$).

A second difficulty comes from the impossibility of inverting equation (1.5) in a mathematically exact way, i.e., of finding the "true" function $\widehat{x}_1$ among other candidate functions, even in the absence of noise. Consider, for example, the FT $\widehat{\widehat{x}}$ of a stable series $\{\widehat{x}_n\}_{\mathbb{Z}}$ such that:

$$\widehat{x}_n = y_n/h_n \quad \text{if} \quad n \in \{1, \ldots, N\} \quad \text{and} \quad h_n \neq 0\,. \tag{1.6}$$

Since this series is only defined for N values at most, there is an infinite number of solutions $\widehat{\widehat{x}}$ that satisfy constraints (1.6), and are equivalent considering the data. The problem is therefore *indeterminate*. In this respect, the periodogram of the data:

$$\Gamma(\nu) \triangleq \frac{1}{N} \left| \sum_{n=1}^{N} y_n \exp\{-2j\pi\nu n\} \right|^2, \qquad \nu \in [0,1]\,, \tag{1.7}$$

calculable by fast discrete FT on a fine, regularly spaced grid, is a particular solution for h_n close to 1 (i.e., $\widehat{h}$ close to a Dirac). It is obtained by extending $\widehat{x}_n = y_n$ with zeros on either side of the observation window.

The small number of data points and the spread of the instrument response $\widehat{h}$ give the periodogram very low resolving power (see Figure 1.2a, curve (P_1)). We can try to get around the need to have $\widehat{h}$ by calculating the periodogram associated with y_n/h_n

or, in other words, by making a spectral estimator $\widehat{x}$ from a time series extrapolating y_n/h_n with zeros. It is also worth noting that this is none other than the trivial solution to the problem of finding a series $\{x_n\}_{\mathbb{Z}}$ that is stable and has a minimal norm, and which minimizes the least squares criterion – even reducing it to zero in this case:

$$\sum_{n=1}^{N}(y_n - h_n\, x_n)^2\,.$$

The result is disappointing (see Figure 1.2a, curve (P_2)). In fact, this is not really surprising as the series y_n/h_n contains aberrant values at its extremities because of the measurement noise and quantification. These error terms, which are amplified by $1/h_n$ when h_n is small, make a contribution to the estimated spectrum that completely masks the peaks of the theoretical spectrum.

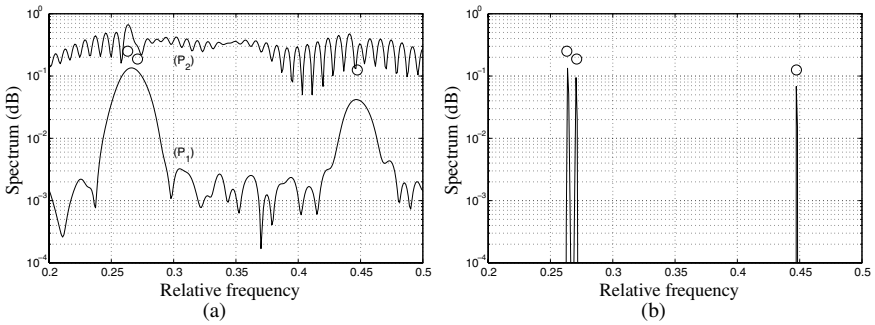

Figure 1.2. (a) *Curve* (P_1) *is the periodogram of the data* y_n *represented in dB; the lack of resolution is a result of the lack of data but also of the spread response of the instrument. Curve* (P_2) *is the periodogram associated with* y_n/h_n*;* (b) *spectral estimate obtained as the minimizer of criterion* (1.8)*, calculated by approximation on a discrete grid of* $1{,}024$ *points, for "well chosen" values of* λ *and* τ

These negative results could lead us to think that the data is too poor to be used. This is not the case, as shown by the spectral estimate whose modulus is represented in Figure 1.2b, and which is obtained as the function $\widehat{x}$ that minimizes the *regularized* criterion:

$$\sum_{n=1}^{N}(y_n - h_n\, x_n)^2 + \lambda \int_0^1 \sqrt{\tau^2 + |\widehat{x}(\nu)|^2}\; d\nu\,, \tag{1.8}$$

where x_n is connected to $\widehat{x}$ by (1.4) for "well chosen" values of *hyperparameters* λ and τ. As the FT $\widehat{x}$ is discretized on $1{,}024$ points, this process is strictly equivalent to extrapolating the series of 128 observed x_n by 896 values that are not necessarily zero, unlike in the periodogram.

This example is typical of the difficulties that arise in the solving of numerous inverse problems [AND 77, BER 88, HER 87]. Certain conventional signal processing

tools prove to be unsuitable whereas others provide qualitatively and quantitatively exploitable solutions. The improvement obtained with this regularized criterion (1.8) is striking and several questions immediately come to mind: why do we need to penalize the least squares criterion in this way? How do we obtain the argument of the minimum of such a criterion? How do we choose the values of the hyperparameters that are part of it? It can be said that the main part of this book is devoted to just that: the construction and use of regularized criteria. However, it is important to understand the nature of the difficulties encountered during inversion before studying the regularized solutions that allow them to be solved.

1.3. Ill-posed problem

The aim of this section is to correct the false impression that the difficulties encountered in solving an inverse problem come from the discrete nature of the data and its finite amount and that, if we had access to a *continuum* of values, i.e., the function $y(t)$ in the example above, everything would be fine. Often unsuspected difficulties are already present at this level. They are proper to problems known as *ill-posed* problems. When the problem is inevitably discretized as in the previous example, some of these difficulties paradoxically disappear, but the problem most often remains *ill-conditioned*.

Hadamard has defined three conditions for a mathematical problem to be *well-posed* [AND 80, HAD 01, NAS 81, TIK 77] (by default, it will be called *ill-posed*):

(a) for each item of data y in a defined class $\mathcal{Y}$, there exists a solution x in a stipulated class $\mathcal{X}$ (*existence*);

(b) the solution is unique in $\mathcal{X}$ (*uniqueness*);

(c) the dependence of x on y is continuous, i.e., when the error δy on data item y tends towards zero, the error δx induced on the solution x also tends towards zero (*continuity*).

The continuity requirement is connected to that of *stability* or *robustness* of the solution (with respect to the errors that inevitably occur on the data). Continuity is, however, a necessary but not a sufficient condition for robustness [COU 62]. A well-posed problem can be *ill-conditioned*, which makes its solution non-robust, as we shall see in section 1.5.

All the traditional problems of mathematical physics, such as the Dirichlet problem for elliptical equations or the Cauchy problem for hyperbolic equations, are well-posed in Hadamard's sense [AND 80]. However, the "inverse" problems obtained from "direct" problems by exchanging the roles of the solution and data are generally *not* well-posed.

The example of section 1.2 clearly comes into this category of ill-posed problems since, with a finite number of discrete data items, a solution $\widehat{x}(\nu)$ exists, but it is not unique. It is interesting to note that this same problem, before discretization, is a special case of the general problem of solving a Fredholm integral equation of the first kind:

$$y(s) = \int k(s,\, r)\, x(r)\, dr\,, \tag{1.9}$$

where $y(s)$, $x(r)$ and $k(s,\, r)$ are replaced by $y(t)$, $\widehat{x}(\nu)$ and $h(t)\, \exp\{2j\pi\nu t\}$ respectively. Later in this book we will find the same type of integral equations for other forms of kernel $k(s,\, r)$, for deconvolution, tomographic reconstruction and Fourier synthesis.

As the data is uncertain or noisy, we cannot hope to solve this equation exactly and the solution needs to be approached from a certain direction. The concept of distance between functions is thus a natural way of evaluating the quality of an approximation, which explains why x and y are often assumed to belong to Hilbert spaces. Problem (1.9) can thus be rewritten as:

$$y = A\, x\,, \qquad x \in \mathcal{X}, \qquad y \in \mathcal{Y}, \tag{1.10}$$

where x and y are now elements of functional spaces of infinite dimension $\mathcal{X}$ and $\mathcal{Y}$, respectively, and where A: $\mathcal{X} \rightarrow \mathcal{Y}$ is the linear operator corresponding to (1.9). The necessary and sufficient conditions for the existence, uniqueness and continuity of the solution can thus be written respectively [NAS 81]:

$$\mathcal{Y} = \operatorname{Im} A\,, \qquad \operatorname{Ker} A = \{0\}\,, \qquad \operatorname{Im} A = \overline{\operatorname{Im} A}\,, \tag{1.11}$$

where $\operatorname{Im} A$ is the *image* of A (i.e., the set of y that are images of an $x \in \mathcal{X}$), $\operatorname{Ker} A$ its *kernel* (i.e., the set of solutions to the equation $A\, x = 0$) and $\overline{\operatorname{Im} A}$ the *closure* of $\operatorname{Im} A$ [BRE 83].

The manner in which conditions (1.11) are stated gives rise to several comments. On the one hand, $\mathcal{Y} = \operatorname{Im} A$ implies $\operatorname{Im} A = \overline{\operatorname{Im} A}$ (a Hilbert space is closed upon itself). In other words, the very existence of a solution to problem (1.9) $\forall\, y \in \mathcal{Y}$ implies the continuity of this solution. *In contrast*, if the existence condition $\mathcal{Y} = \operatorname{Im} A$ is not verified, the continuity condition seems to become pointless; in fact, it applies to *pseudo-solutions*, which will be defined in section 1.4.1 as minimizing the norm $\|A\, x - y\|_{\mathcal{Y}}$ (without systematically reducing it to zero).

1.3.1. *Case of discrete data*

When the data is discrete, y is a vector of dimension N in a Euclidian space. Ignoring errors on the data, a linear inverse problem with discrete data can be stated

as follows. Given a set $\{F_n(x)\}_{n=1}^N$ of linear functionals defined on $\mathcal{X}$ and a set $\{y_n\}_{n=1}^N$ of numbers, find a function $x \in \mathcal{X}$ such that:

$$y_n = F_n(x)\,, \qquad n = 1, \ldots, N.$$

In particular, when functionals F_n are continuous on $\mathcal{X}$, Riesz theorem [BRE 83] states that functions $\psi_1, \ldots, \psi_N$ exist such that:

$$F_n(x) = \langle x,\, \psi_n \rangle_{\mathcal{X}}\,,$$

where the notation $\langle \cdot\,, \cdot \rangle_{\mathcal{X}}$ designates the scalar product used in space $\mathcal{X}$. The example of equation (1.9) takes this form when $y(s)$ is measured on a finite number of points $s_1, \ldots, s_N$, and $\mathcal{X}$ is an L^2 space. In this case we have:

$$\psi_n(r) = k(s_n, r)\,.$$

This problem is a particular case of that of equation (1.10) if we define an operator A of $\mathcal{X}$ in $\mathcal{Y}$ by the relation:

$$(A\,x)_n = \langle x, \psi_n \rangle_{\mathcal{X}} \qquad n = 1, \ldots, N.$$

Operator A is not injective: $\mathrm{Ker}\,A$ is the closed subspace of infinite dimension of all the functions x orthogonal to the subspace engendered by the functions ψ_n. Conversely, the image of A, $\mathrm{Im}\,A$ is closed: $\mathrm{Im}\,A$ is simply $\mathcal{Y}$ when the functions ψ_n are linearly independent; otherwise it is a subspace of dimension $N' < N$. We thus see clearly why the example of section 1.2 is an ill-posed problem: the difficulty does not lie in a lack of continuity but in a lack of uniqueness.

1.3.2. *Continuous case*

Let us now assume that x and y belong to the same Hilbert space and that k is square integrable, a condition fulfilled by many imaging systems – it would be the situation if our example of section 1.2 was modified so that the function $\widehat{y}(\nu) = \widehat{h} \star \widehat{x}_1(\nu)$ was continuously observed. The direct problem is thus well-posed: a small error δx on the data entails a small error δy on the solution. This condition is not, however, fulfilled in the corresponding inverse problem, where it is object x that must be calculated from response y: $x = A^{-1}\,y$. In fact, when kernel k is square integrable – which would be the case for a Gaussian kernel in our example – the Riesz-Fréchet theorem indicates that operator A is bounded and *compact* [BRE 83]. However, the image of a compact operator is not closed (except in the degenerate case where its dimension is finite). This signifies that the inverse operator A^{-1} is not bounded, or stable, its image is not closed and the third of Hadamard's conditions is not satisfied for the inverse problem [NAS 81].

To get a better grasp of these abstract ideas, it is handy to use the spectral properties of compact operators in Hilbert spaces. The most remarkable property of these operators is that they can be *decomposed into singular values*, like matrices (the famous *singular value decomposition*, or SVD). The singular system of a compact operator is defined as the set of solutions of the coupled equations:

$$A\,u_n = \sigma_n\,v_n \quad \text{and} \quad A^*\,v_n = \sigma_n\,u_n\,, \tag{1.12}$$

where the *singular values* σ_n are positive numbers, where the *singular functions* u_n and v_n are elements of $\mathcal{X}$ and $\mathcal{Y}$respectively, and where A^* is the *adjoint operator* of A, which exists since A is continuous and therefore such that: $\langle A\,x, y\rangle_{\mathcal{Y}} = \langle x, A^*\,y\rangle_{\mathcal{X}}$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ [1].

When A is compact, it always possesses a singular system $\{u_n, v_n;\ \sigma_n\}$ with the following properties [NAS 81]:

– σ_n being ordered and counted with their multiplicity (which is finite): $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq \ldots 0$, σ_n tends towards 0 when $n \to \infty$ and either the limit is reached for $n = n_0$ (in which case operator A is degenerate), or it is not reached for any finite value of n;

– functions u_n form an orthonormal basis of $(\mathrm{Ker}\,A)^\perp$, the orthogonal complement of $\mathrm{Ker}\,A$ in the decomposition: $\mathcal{X} = \mathrm{Ker}\,A \oplus (\mathrm{Ker}\,A)^\perp$ and the functions v_n form an orthonormal basis of $(\mathrm{Ker}\,(A^*))^\perp$, i.e., $\overline{\mathrm{Im}\,A}$, orthogonal complement of $\mathrm{Ker}\,(A^*)$ in the decomposition $\mathcal{Y} = \mathrm{Ker}\,(A^*) \oplus (\mathrm{Ker}\,(A^*))^\perp$.

Let $E \subseteq \mathbb{N}$ be the set of indices n such that $\sigma_n \neq 0$. The Picard criterion [NAS 81] ensures that a function $y \in \mathcal{Y}$ is in $\mathrm{Im}\,A$ if and only if:

$$y \in (\mathrm{Ker}\,(A^*))^\perp \quad \text{and} \quad \sum_{n\in E} \sigma_n^{-2}\ \langle y, v_n\rangle^2 < +\infty\,. \tag{1.13}$$

For the second condition (1.13) to be satisfied, it is necessary, when operator A is not degenerate ($E \equiv \mathbb{N}$), for the components $\langle y, u_n\rangle$ of the development of image y on the set of eigenfunctions $\{v_n\}$ to tend towards zero faster than the eigenvalues σ_n^2 when $n \to \infty$. This strict condition has no reason to be satisfied by an arbitrary function of $(\mathrm{Ker}\,(A^*))^\perp$. Note, however, that it is naturally satisfied if $y = A\,x$ is the perfect image resulting from an object x of finite energy. The solution is thus written:

$$x = \sum_{n\in E} \sigma_n^{-1}\ \langle y, v_n\rangle\ u_n\,. \tag{1.14}$$

1. Note that the self-adjoint operator A^*A, which appears in the symmetrized problem $A^*\,y = A^*A\,x$, verifies: $A^*A\,v_n = \sigma_n^2\,v_n$. It is thus defined as non-negative since its eigenvalues are σ_n^2 (which are also those of $A\,A^*$). This property will be used in section 2.1.1.

However, this solution, when it exists, is *unstable*: a small additive perturbation $\delta y = \varepsilon\, v_N$, for example, on the perfect data y leads to a perturbation δx on the solution calculated with the data $y + \delta y$:

$$\delta x = \sigma_N^{-1}\ \langle \delta y, v_N \rangle\ u_N = \sigma_N^{-1}\, \varepsilon\, u_N\,. \tag{1.15}$$

The ratio $\|\delta x\| \,/\, \|\delta y\|$ equals σ_N^{-1}, which can be arbitrarily large. The inverse linear operator $A^{-1} : \mathcal{Y} \to \mathcal{X}$, defined by (1.14), is thus not *bounded* as it is not possible to find a constant C such that, for all $y \in \mathcal{Y}$, we have $\|A^{-1}y\|_{\mathcal{X}} \leq C\, \|y\|_{\mathcal{Y}}$, which is a necessary and sufficient condition for A^{-1} to be *continuous* [BRE 83]. The ill-posed nature of the problem stems this time from the lack of continuity and not from the lack of uniqueness.

The need for a deeper understanding of these problems that are not mathematically well-posed but are of great interest in engineering sciences is at the origin of two recent branches of analysis: *generalized inversion* theory [NAS 76], which is summarized below, and *regularization* theory, which will be the subject of the next chapter.

1.4. Generalized inversion

Let us suppose that the equation $A\,x = 0$ has non-trivial solutions. The set $\mathrm{Ker}\, A \neq \{0\}$ of these solutions is a closed subspace of $\mathcal{X}$. It is the set of "invisible objects" as they produce an image y that is zero. Let us also suppose that $\mathrm{Im}\, A$ is a closed subspace of $\mathcal{Y}$. An example is provided by the integral operator corresponding to an ideal low-pass filter of cut-off pulsation Ω [BER 87]:

$$(A\,x)(r) = \int_{-\infty}^{+\infty} \frac{\sin \Omega(r - r')}{\pi(r - r')}\, x(r')\, dr'\,. \tag{1.16}$$

If we choose $\mathcal{X} = \mathcal{Y} = L^2_{\mathbb{R}}$, the kernel is the set of all the functions x whose FT is zero in the band $[-\Omega, +\Omega]$, while the image of A is the set of functions having a limited band in the same interval, which is a closed subspace of $L^2_{\mathbb{R}}$.

A means of re-establishing the existence and the uniqueness of the solution in the above conditions is to redefine both the space $\mathcal{X}$ of the solutions and the space $\mathcal{Y}$ of the data. If we choose a new space $\mathcal{X}'$ which is the set of all the functions orthogonal to $\mathrm{Ker}\, A$ (in the case of equation (1.16), $\mathcal{X}'$ is the set of functions with summable squares and band limited to the interval $[-\Omega, +\Omega]$), and if y is restrained to a new data space $\mathcal{Y}' = \mathrm{Im}\, A$ (which is, once again, in the case of equation (1.16), the set of functions with summable squares and band limited to the interval $[-\Omega, +\Omega]$), thus, for any $y \in \mathcal{Y}'$, there exists a unique $x \in \mathcal{X}'$ such that $A\,x = y$ (in our example (1.16), the solution is even trivial: $x = y$) and the new problem is thus well-posed.

It is often possible to choose the spaces $\mathcal{X}$ and $\mathcal{Y}$ so that the problem becomes well-posed but the practical interest of the choice is limited because it is generally the

intended application that imposes the appropriate spaces. Another means that could be envisaged is to change the notion of a solution itself.

1.4.1. *Pseudo-solutions*

Let us first consider the case where A is injective ($\mathrm{Ker}\, A = \{0\}$) but not surjective ($\mathrm{Im}\, A \neq \mathcal{Y}$). The set of functions x that are solutions of the variational problem:

$$x \in \mathcal{X} \text{ minimizes } \|A\,x - y\|_{\mathcal{Y}}\ , \tag{1.17}$$

where $\|\cdot\|_{\mathcal{Y}}$ designates the norm in $\mathcal{Y}$, are called pseudosolutions or least squares solutions of the problem (1.10). If $\mathrm{Im}\, A$ is closed, (1.17) always has a solution, but it is not unique if the kernel $\mathrm{Ker}\, A$ is not trivial. When it is, as we assume here, it can be said that the well-posed character has been restored by reformulating the problem in the form (1.17).

By making the first variation of the function minimized in (1.17) zero, we obtain Euler's equation:

$$A^*A\,x = A^*y\,, \tag{1.18}$$

which brings in the self-adjoint operator A^*A, the eigensystem of which can be deduced from the singular system of A.

1.4.2. *Generalized solutions*

Let us now consider the case where the uniqueness condition is not satisfied ($\mathrm{Ker}\, A \neq \{0\}$, the problem is indeterminate). The set of solutions of (1.18) being a convex, closed subset of $\mathcal{X}$, it contains a single element with a minimal norm, noted $x^\dagger$ or $\widehat{x}^{\text{GI}}$ and called the *generalized solution* of (1.10). As $x^\dagger$ is orthogonal to $\mathrm{Ker}\, A$, this way of defining the solution is equivalent to choosing $\mathcal{X}' = (\mathrm{Ker}\, A)^\perp$. In other words, the generalized solution is a least squares solution having the minimal norm among these solutions. As there is a single $x^\dagger$ for every $y \in \mathcal{Y}$, a linear application $A^\dagger$ of $\mathcal{Y}$ in $\mathcal{X}$ is defined by:

$$A^\dagger y = x^\dagger = \widehat{x}^{\text{GI}}\,. \tag{1.19}$$

The operator $A^\dagger$ is called the *generalized inverse* of A and is continuous [NAS 76].

1.4.3. *Example*

To illustrate the idea of generalized inversion, let us go back to our example of section 1.2 and, first of all, neglect the weighting $h(t)$. To impose a unique solution in

the class of possible solutions prolonging the series of known x_n, we can choose the generalized inverse solution of the initial problem (1.10):

$$\widehat{\widehat{x}}^{\text{GI}}(\nu) = \underset{\widehat{x}_1 \in L^2_{\mathbb{C}}[0,1]}{\arg\min} \int_0^1 |\widehat{x}_1(\nu)|^2 \, d\nu \;\text{ subject to (s.t.) }\; x_n = y_n,\; n = 1, \ldots, N.$$

The Plancherel and Parseval relations show us that this is equivalent to finding coefficients:

$$\widehat{x}_n = \int_0^1 \widehat{\widehat{x}}(\nu) \, \exp\{2j\pi n\nu\} \, d\nu\,, \quad n \in \mathbb{Z}\,,$$

such that:

$$\widehat{\boldsymbol{x}} = \underset{x \in \ell^2_{\mathbb{C}}}{\arg\min} \sum_{n \in \mathbb{Z}} |x_n|^2, \;\text{ s.t. }\; x_n = y_n\,, \;\; n = 1, \ldots, N.$$

The solution is trivial since the problem is separable:

$$\widehat{x}_n = \begin{cases} y_n & \text{if} \quad n \in \{1,\, 2, \ldots,\, N\}, \\ 0 & \text{otherwise,} \end{cases} \quad \Longrightarrow \; \widehat{\widehat{x}}^{\text{GI}}(\nu) = \sum_{n=1}^{N} y_n \, e^{-2j\pi n\nu}, \quad (1.20)$$

whose squared modulus, with just the difference of a coefficient, gives the *Schuster periodogram* of equation (1.7). The case of weighting by $h(t)$ is treated in the same way, by replacing y_n by y_n/h_n. It can thus be seen that the periodogram is the generalized inverse solution of a spectral analysis problem that is ill-posed because the number of data points is finite.

1.5. Discretization and conditioning

A first description of a direct problem often brings in functions of real variables (time, frequency, space variables, etc.), representing the physical quantities involved: quantities accessible for measurement and quantities of interest that are unknown. The analysis of the problem at this level of description has provided an explanation for the difficulties that arise during inversion, by situating us in functional spaces of infinite dimension. We have thus seen that, in the case of a direct problem described by an integral equation of the first kind, the inversion is often an ill-posed problem as it is unstable.

This analysis is, however, insufficient. The available experimental data are almost always composed of measurements of physical quantities accessible at a necessarily finite number of points in the domain of definition of their variables. They are thus naturally discrete and we group them together in the vector $\boldsymbol{y}$ as we did in section 1.4 above. However, the unknown object is also discretized, either right from the start, or during the process of solution (as in the example of the periodogram above), by

decomposition over a finite number of functions. If these are basic elements of the space to which the object belongs, the decomposition is necessarily truncated. In imaging, for example, in the vast majority of cases, pixel indicators or cardinal sines are used as basic functions, according to whether the object is implicitly assumed to have a limited support or a limited spectrum. Basic wavelets or wavelet packets are also coming into use [STA 02, KAL 03]. The starting point is thus composed of a *model parametrized* by the vector $\boldsymbol{x}$ of the decomposition coefficients, in other words by a set of exclusive hypotheses, each of which is indexed by the value of the coefficients. This hypothesis space is thus the set of possible values of these unknown parameters, $\mathcal{H} = \{x_i\}$. The choice of these basic functions obviously forms part of the inversion problem, even if it is not often touched upon.

In the discrete case (or more exactly the "discrete-discrete" case), the problem changes noticeably as $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to spaces of finite dimensions and the linear operator A becomes a matrix $\mathbf{A}$. Equation (1.10) has a unique solution with minimal norm $\widehat{\boldsymbol{x}}^{\text{GI}} = \mathbf{A}^{\dagger}\boldsymbol{y}$ which depends continuously on $\boldsymbol{y}$ since the generalized inverse $\mathbf{A}^{\dagger}$ is then always bounded [NAS 76]. The problem is thus always well-posed in Hadamard's sense. However, even in this framework, the inversion problem still has an unstable nature, this time from a numerical point of view. The spectral decomposition (1.15) is still valid, the only difference being that the number of singular values of matrix $\mathbf{A}$ is now finite. These singular values can rarely be calculated explicitly [KLE 80]. From this point of view, the example in section 1.2 is not representative because, if we choose to decompose the solution over $M > N$ complex exponentials of the Fourier basis:

$$\widehat{x}(\nu) = \sum_{m=1}^{M} x_m \, \exp\left\{-2j\pi m\nu\right\},$$

model (1.2) can be written $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{b}$ in a matrix-vector notation, where $\mathbf{A}$ is a rectangular $N \times M$ matrix composed of the diagonal matrix $\operatorname{diag}\{h_n\}$, juxtaposed with the zero matrix of size $N \times (M - N)$. Its singular values are thus $\sigma_n = h_n$ for $n = 1, 2, \ldots, N$ and $\sigma_n = 0$ otherwise. Even if we exclude all the zero singular values, as by using $\mathbf{A}^{\dagger}$, there are always some singular values close to zero with the weighting $h(t)$ of our example. Matrix $\mathbf{A}$ is thus *ill-conditioned*. The coefficients $\sigma_n^{-1} \langle \delta\boldsymbol{y}, \boldsymbol{u}_n \rangle$ in equation (1.15) become very large for the σ_n that are close to zero, even if $\delta\boldsymbol{y}$ is small.

Generally speaking, whether we have the discrete case or not, let us assume that $\operatorname{Im} A$ is closed so that the generalized inverse $A^{\dagger}$ exists $\forall\, y \in \mathcal{Y}$ (and is continuous). Let us designate an error on the data y as δy and the error induced on the generalized inverse solution $x^{\dagger}$ as $\delta x^{\dagger}$. The linearity of (1.19) leads to $\delta x^{\dagger} = A^{\dagger}\delta y$, which implies

$$\|\delta x^{\dagger}\|_{\mathcal{X}} \leq \|A^{\dagger}\| \; \|\delta y\|_{\mathcal{Y}} \;,$$

where $\|A^\dagger\|$ designates the norm of the continuous operator $A^\dagger$, that is to say the quantity: $\sup_{y\in\mathcal{Y}} \|A^\dagger y\|_\mathcal{X} / \|y\|_\mathcal{Y}$ [BRE 83]. In a similar way, (1.10) implies:

$$\|y\|_\mathcal{Y} \leq \|A\| \ \|x^\dagger\|_\mathcal{X} \,,$$

where $\|A\| = \sup_{x\in\mathcal{X}} \|A\,x\|_\mathcal{Y} \,/\, \|x\|_\mathcal{X}$. By combining these two relations, we obtain the inequality:

$$\frac{\|\delta x^\dagger\|_\mathcal{X}}{\|x^\dagger\|_\mathcal{X}} \ \leq \ \|A\| \ \|A^\dagger\| \ \frac{\|\delta y\|_\mathcal{Y}}{\|y\|_\mathcal{Y}} \,. \tag{1.21}$$

It is important to note that this inequality is precise in a certain sense. When A is a matrix of dimensions $(N \times M)$ or corresponds to an inverse problem with discrete data, the inequality can become an equality for certain $(y,\, \delta y)$ pairs. When A is an operator on spaces of infinite dimension, it can only be established that the left hand side of inequality (1.21) can be arbitrarily close to the right hand side. The quantity:

$$c = \|A\| \ \|A^\dagger\| \geq 1 \tag{1.22}$$

is called the *condition number* of the problem. When c is close to one, the problem is said to be *well-conditioned*, whereas when it is considerably larger than one, the problem is said to be *ill-conditioned*.

In practice, it is useful to have an estimate of the condition number, which gives an idea of the numerical stability of the problem. When $A = \mathbf{A}$ is a matrix of dimensions $(N \times M)$, $\|\mathbf{A}\|$ is the square root of the largest of the eigenvalues of the positive semidefinite symmetric matrix $\mathbf{A}^*\mathbf{A}$, of dimensions $(M \times M)$ (the positive eigenvalues of this matrix coincide with those of the matrix $\mathbf{A}\mathbf{A}^*$) and $\|\mathbf{A}^\dagger\|$ is the inverse of the square root of the smallest of these eigenvalues:

$$c = \sqrt{\lambda_{\max}/\lambda_{\min}} \,.$$

In our example in section 1.2, we obtain $c = |h|_{\max}/|h|_{\min}$ and we understand why the weighting by $h(t)$ can degrade the conditioning of the generalized inversion problem which is otherwise well-posed.

1.6. Conclusion

To sum up the above, when we have a simple situation where we are dealing with a direct, linear problem in an infinite dimension, bringing in an operator $A : \mathcal{X} \to \mathcal{Y}$ defined in the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, we have three main situations:

– if A is continuous and injective (the only solution to the equation $A\,x = 0$ is the trivial solution $x = 0$, thus $\mathrm{Ker}\, A = \{0\}$) and its image is closed and given by $\mathrm{Im}\, A = \mathcal{Y}$, the inverse problem is well-posed, since the inverse operator is continuous;

– if A is not injective, but $\mathrm{Im}\, A$ is closed, then, if we look for a pseudosolution, the inverse problem becomes well-posed in as far as the generalized inverse is continuous;

– if the image $\mathrm{Im}\, A$ is not closed, using a pseudo-solution cannot, in itself, guarantee the existence and the continuity of the inverse solution.

When we are dealing with a linear operator defined in spaces of finite dimension $\mathbb{R}^N$ and $\mathbb{R}^M$ and of the type $A\colon \mathbb{R}^M \to \mathbb{R}^N$, we again have three main situations:

– if p is the rank of the matrix associated with the operator A and if $p = N = M$, then A is bijective. A solution always exists, it is unique, and the inverse problem is well defined;

– if $p < M$, then the uniqueness is not certain but can be established by considering a generalized inversion;

– if $p < N$, then the existence is not certain for any given data but can be ensured by again considering a generalized inversion.

To conclude this chapter, we see that an inverse problem is often ill-posed or ill-conditioned, and that generalized inversion does not, in general, provide a satisfactory solution. In the next chapter we will see that another development of modern analysis, regularization, allows us to get around these difficulties and gives a generic framework for inversion.

1.7. Bibliography

[AND 77] ANDREWS H. C., HUNT B. R., *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[AND 80] ANDERSSEN R. S., DE HOOG F. R., LUKAS M. A., *The Application and Numerical Solution of Integral Equations*, Sijthoff and Noordhoff, Alplen aan den Rijn, 1980.

[BER 87] BERTERO M., POGGIO T., TORRE V., Ill-posed Problems in Early Vision, Memo 924, MIT, May 1987.

[BER 88] BERTERO M., DE MOL C., PIKE E. R., "Linear inverse problems with discrete data: II. Stability and regularization", *Inverse Problems*, vol. 4, p. 573-594, 1988.

[BRE 83] BREZIS H., *Analyse fonctionnelle : théorie et applications*, Masson, Paris, 1983.

[COU 62] COURANT R., HILBERT D., *Methods of Mathematical Physics*, Interscience, London, 1962.

[FOS 61] FOSTER M., "An application of the Wiener-Kolmogorov smoothing theory to matrix inversion", *J. Soc. Indust. Appl. Math.*, vol. 9, p. 387-392, 1961.

[FRA 70] FRANKLIN J. N., "Well-posed stochastic extensions of ill-posed linear problems", *J. Math. Anal. Appl.*, vol. 31, p. 682-716, 1970.

[GEM 84] GEMAN S., GEMAN D., "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, num. 6, p. 721-741, Nov. 1984.

[HAD 01] HADAMARD J., "Sur les problèmes aux dérivées partielles et leur signification physique", *Princeton University Bull.*, vol. 13, 1901.

[HER 87] HERMAN G. T., TUY H. K., LANGENBERG K. J., SABATIER P. C., *Basic Methods of Tomography and Inverse Problems*, Adam Hilger, Bristol, UK, 1987.

[KAK 88] KAK A. C., SLANEY M., *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, NY, 1988.

[KAL 03] KALIFA J., MALLAT S., ROUGÉ B., "Deconvolution by thresholding in mirror wavelet bases", *IEEE Trans. Image Processing*, vol. 12, num. 4, p. 446-457, Apr. 2003.

[KLE 80] KLEMA V. C., LAUB A. J., "The singular value decomposition: its computation and some applications", *IEEE Trans. Automat. Contr.*, vol. AC-25, p. 164-176, 1980.

[MAR 87] MARROQUIN J. L., MITTER S. K., POGGIO T. A., "Probabilistic solution of ill-posed problems in computational vision", *J. Amer. Stat. Assoc.*, vol. 82, p. 76-89, 1987.

[NAS 76] NASHED M. Z., *Generalized Inverses and Applications*, Academic Press, New York, 1976.

[NAS 81] NASHED M. Z., "Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory", *IEEE Trans. Ant. Propag.*, vol. 29, p. 220-231, 1981.

[PHI 62] PHILLIPS D. L., "A technique for the numerical solution of certain integral equation of the first kind", *J. Ass. Comput. Mach.*, vol. 9, p. 84-97, 1962.

[SAB 78] SABATIER P. C., "Introduction to applied inverse problems", in SABATIER P. C. (Ed.), *Applied Inverse Problems*, p. 2-26, Springer Verlag, Berlin, Germany, 1978.

[STA 02] STARK J.-L., PANTIN E., MURTAGH F., "Deconvolution in astronomy: a review", *Publ. Astr. Soc. Pac.*, vol. 114, p. 1051-1069, 2002.

[TIK 63] TIKHONOV A., "Regularization of incorrectly posed problems", *Soviet. Math. Dokl.*, vol. 4, p. 1624-1627, 1963.

[TIK 77] TIKHONOV A., ARSENIN V., *Solutions of Ill-Posed Problems*, Winston, Washington, DC, 1977.

[TWO 62] TWOMEY S., "On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature", *J. Assoc. Comp. Mach.*, vol. 10, p. 97-101, 1962.

Chapter 2

Main Approaches to the Regularization of Ill-posed Problems

In the previous chapter, we saw that, when the image $\mathrm{Im}\, A$ of a linear operator we want to invert is not closed, then the inverse A^{-1}, or the generalized inverse $A^{\dagger}$, is not defined everywhere in the data space $\mathcal{Y}$ and is not continuous. This is the case, for example, of compact, non-degenerate (or non-finite rank) operators and it is easy to see that the condition number of the problem is infinite. Suitable solving techniques are thus required.

We also saw that, in a finite dimension, the inverse or the generalized inverse is always continuous. In consequence, the use of a generalized inversion is sufficient to guarantee that the problem is well posed in this case. However, it must not be forgotten that a problem that is well posed but severely ill-conditioned behaves in practice like an ill-posed problem and has to be treated with the same regularization methods, which we present below.

2.1. Regularization

In a finite or infinite dimension, a *regularizer* of equation (1.10) $y = Ax$ is a family of operators $\{R_\alpha;\ \alpha \in \Lambda\}$ such that [NAS 81, TIK 63]:

$$\begin{cases} \forall\, \alpha \in \Lambda, & R_\alpha \text{ is a continuous operator of } \mathcal{Y} \text{ in } \mathcal{X}; \\ \forall\, y \in \mathrm{Im}\, A, & \lim_{\alpha \to 0} R_\alpha\, y = A^{\dagger} y. \end{cases} \tag{2.1}$$

Chapter written by Guy DEMOMENT and Jérôme IDIER.

In other words, since the inverse operator A^{-1} does not have the required continuity or stability properties, we construct a family of continuous operators, indexed by a regulating parameter α (called the *regularization coefficient*) and including $A^\dagger$ as a limit case. Applied to perfect data y, R_α gives an approximation of $x^\dagger$ that is all the better as $\alpha \to 0$. However, when R_α is applied to data $y_\varepsilon = Ax + b$ that inevitably contain noise, b, we obtain an *approximate* solution $x_\varepsilon = R_\alpha\, y_\varepsilon$ and we have:

$$R_\alpha\, y_\varepsilon = R_\alpha\, y + R_\alpha\, b\,. \tag{2.2}$$

The second term diverges when $\alpha \to 0$. It follows that a trade-off has to be made between two opposing terms, the approximation error (first term) and the error due to noise (second term). This can be done, within a given family of operators R_α, by adjusting the value of the coefficient of regularization α.

Most of the methods that have been put forward for solving and stabilizing ill-posed problems in the past 30 years fall into this general scheme in one way or another. They can be divided into two broad families: those that proceed by dimensionality control – Λ is thus a discrete set – and those that work by minimization of a composite criterion or by optimization under constraint – Λ is thus $\mathbb{R}_+$. In what follows, we will mainly concern ourselves with the second family of regularization methods.

2.1.1. *Dimensionality control*

In the case of an ill-posed or ill-conditioned problem, the methods of regularization by dimensionality control get around the difficulty in two ways:

– by minimizing the criterion $\|y - Ax\|$ (or, more generally, $\mathcal{G}(y - Ax)$) in a subspace of reduced dimension, after an appropriate change of basis if necessary;

– by minimizing the criterion $\mathcal{G}(y - Ax)$ in the space initially chosen but by an iterative method in which the number of iterations is limited.

2.1.1.1. *Truncated singular value decomposition*

A typical example of methods of the first family can be found by examining equation (1.15): to suppress the ill-conditioned nature of the problem, we just truncate the development, keeping the components corresponding to singular values that are large enough for error terms of the form $\sigma_n^{-1} \langle \delta y, v_n \rangle\, u_n$ to remain small. This is *truncated singular value decomposition*, or TSVD [AND 77, NAS 81]. It is very effective for ensuring numerical stability. However, the problem arises of choosing the truncation order, which plays the role, here, of the inverse of a regularization coefficient. However, the main failing of this approach is that we give up the possibility of re-establishing the spectral components that have been too degraded by the imaging device. As for the definition of the Rayleigh resolution criterion in optics, we use no information about the object sought other than the fact that its energy is finite although,

most of the time, we know that it is, for example, positive, or that it contains regions of smooth spatial variation separated by sharp boundaries, or that it has bounded values, or bounded support, etc. If we want to go beyond Rayleigh resolution, it is indispensable to be able to take this type of prior information into account.

2.1.1.2. *Change of discretization*

In truncated singular value decomposition, it is the imaging device that more or less imposes the discretization through singular functions of the corresponding operator. However, we can also avoid the difficulties raised by the poor conditioning of matrix $\mathbf{A}$, as a consequence of the object's being descretized on a Cartesian grid for example, by choosing a *parsimonious parameterization* of the object better suited to its prior properties. This is what wavelet-based decomposition methods [STA 02, KAL 03] do, for example. The principle remains the same: *thresholding* is applied to the coefficients of the decomposition so as to eliminate the subspace dominated by the noise components.

This mode of discretization solves the problem of stability or poor conditioning analyzed above but, even so, does not always provide a satisfactory solution. Everything depends on the decomposition that is chosen.

2.1.1.3. *Iterative methods*

A very popular family of methods is made up of *iterative methods* of the form:

$$x^{(n+1)} = x^{(n)} + \alpha \left(y - A x^{(n)} \right), \qquad n = 0, 1, \ldots \tag{2.3}$$

where $0 < \alpha < 2/\left\| A \right\|$ (Bialy's method [BIA 59]). If A is a non-negative, bounded, linear operator (i.e., $\langle Ax, x \rangle \geq 0$, $\forall x \in \mathcal{X}$) and if $y = Ax$ has at least one solution, then the series of $x^{(n)}$ converges and:

$$\lim_{n \to \infty} x^{(n)} = P\, x^{(0)} + \widehat{x}^{\text{GI}},$$

where P is the orthogonal projection operator on $\mathrm{Ker}\, A$ and $\widehat{x}^{\text{GI}}$ the generalized inverse solution. In fact, this method looks for the fixed point of the operator $G : Gx = \alpha\, y + (I - \alpha\, A)\, x$, but if A is compact and $\mathcal{X}$ is of infinite dimension, then $I - \alpha\, A$ is not a contraction and the method diverges. Moreover, we have also seen that, even in finite dimensions, the generalized inverse solution is often dominated by the noise.

The non-negativity condition excludes a lot of operators but the method can be applied to solve the normal equation $A^* y = A^* A x$ since $A^* A$ is a non-negative operator. We thus obtain Landweber's method [LAN 51]:

$$x^{(n+1)} = x^{(n)} + \alpha\, A^* \left(y - A x^{(n)} \right), \qquad n = 0, 1, \ldots \tag{2.4}$$

with $0 < \alpha < 2/\left\| A^* A \right\|$. The well known Gerchberg-Saxton-Papoulis-VanCittert [BUR 31] method for extrapolating a limited-spectrum signal is a special case of the

Bialy-Landweber method. It is in this same category of iterative methods that we can place Lucy's method [LUC 74], which is very popular in astronomy.

All these methods can provide an acceptable solution only on the condition that the number of iterations is limited (which plays the role of the inverse of a regularization coefficient) [DIA 70]. This is often done empirically, as the initial framework does not take observation noise into account and a theory regulating the number of iterations so as to limit noise amplification cannot but be heteronomous [LUC 94]. This explains why the rest of this book will focus on regularization methods of the second family, which operate by minimization under constraint and are, from this point of view, more autonomous.

2.1.2. *Minimization of a composite criterion*

The principal characteristic of the regularization methods of this second large family is to require the solution to be a trade-off between fidelity to the measured data and fidelity to the prior information [TIT 85]. This trade-off is reached using a single optimality criterion. The approach can be interpreted as follows.

The least squares solutions to equation (1.17) minimize the energy of the discrepancy between the model Ax and the data y. In this sense, they achieve the greatest fidelity to the data. However, when the observation noise is broadband, relation (1.15) shows that the high spatial frequency components of the restored or reconstructed object have large amplitudes because of noise amplification. The least squares solutions thus prove unacceptable because we expect the real object to have markedly smoother spatial variations. We therefore need to introduce a little infidelity to the data to obtain a solution that is smoother than the least squares solution and closer to the idea that we have *a priori*. A widely accepted means of doing this is by the minimization of a composite criterion [NAS 81, TIK 63, TIK 77]. The basic idea is to give up any hope of reaching the exact solution from imperfect data, to consider as *admissible* any solution for which Ax is not far from y, and to look among the admissible solutions to find the one that can be considered as the physically most reasonable, i.e., compatible with certain prior information. This is usually done by finding a solution x_α that minimizes a criterion of the form:

$$\mathcal{J}(x) = \mathcal{G}(y - Ax) + \alpha\,\mathcal{F}(x)\,, \qquad 0 < \alpha < +\infty\,, \tag{2.5}$$

specifically designed so that:

– the solution is faithful to the data up to a certain point (first term of the criterion);

– certain desirable properties that sum up our prior knowledge about the solution are reinforced (second term).

The choice of the functionals $\mathcal{F}$ and $\mathcal{G}$ is qualitative and determines how the regularization is carried out. Conversely, the choice of α, which is the coefficient

of regularization here, is quantitative and allows the compromise between the two sources of information to be adjusted. Perfect fidelity to the data is obtained with $\alpha = 0$, while perfect fidelity to the prior information is obtained if $\alpha = \infty$.

One of the most widely studied regularization methods is obtained by minimizing the functional:

$$\mathcal{J}(x) = \|y - Ax\|_{\mathcal{Y}}^2 + \alpha \, \|Cx\|_{\mathcal{X}}^2 \, , \tag{2.6}$$

where C is a constraint operator [NAS 76]. The existence of a solution is ensured when C is bounded with $\mathrm{Im}\, C$, for example, but that excludes the very interesting case of a differential operator, as in Tikhonov's seminal article [TIK 63]:

$$\|Cx\|_{\mathcal{X}}^2 = \sum_{p=0}^{P} \int c_p(r) \left| x^{(p)}(r) \right|^2 dr \, ,$$

where the weighting functions $c_p(r)$ are strictly positive and $x^{(p)}$ designates the pth order derivative of x. The corresponding regularizer can be written:

$$R_\alpha = (A^* A + \alpha \, C^* C)^{-1} A^* . \tag{2.7}$$

In this case, $x_\alpha = R_\alpha \, y$ exists and is unique when the domain of C is dense in $\mathcal{X}$ and the equations $Ax = 0$ and $Cx = 0$ only have in common the trivial solution $x = 0$. This solution takes a very simple form when A is compact and $C = I$, the identity operator in $\mathcal{X}$. By using the singular value decomposition of A of section 1.3, we obtain:

$$x_\alpha = \sum_{n \in E} \frac{\sigma_n}{\sigma_n + \alpha} \, \frac{1}{\sigma_n} \, \langle y, v_n \rangle \, u_n \, . \tag{2.8}$$

It is thus essentially a "*filtered*" version of the non-regularized solution (1.14), or generalized inverse, of equation (1.10). We will often find this idea of *linear filtering* later, associated with the oldest regularization methods, but, for the moment, we will concern ourselves mainly with discrete problems in finite dimensions.

In the discrete case, the literature on the subject is dominated by a few functionals. Below are the ones most frequently found [TIT 85].

2.1.2.1. *Euclidian distances*

The squared Euclidian distance between two objects $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is defined by:

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\mathbf{P}}^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \mathbf{P} (\boldsymbol{x}_1 - \boldsymbol{x}_2) \, ,$$

where $\mathbf{P}$ is a symmetric positive semi-definite matrix, chosen to express certain desirable characteristics of the proximity measurement. Such a squared distance is the

habitual choice for $\mathcal{G}$ in the case where the noise $\boldsymbol{b}$ is assumed to be zero-mean, Gaussian, independent of $\boldsymbol{x}$, and of probability density:

$$p(\boldsymbol{b}) \propto \exp\left\{-\frac{1}{2}\,\boldsymbol{b}^T\mathbf{P}\,\boldsymbol{b}\right\}\,, \tag{2.9}$$

i.e., of covariance matrix $\mathbf{P}^{-1}$. Such a distance is also very often used for $\mathcal{F}$ in order to penalize objects $\boldsymbol{x}$ of large amplitude.

Applied to the interferometry example of Chapter 1 (which, we recall, is continuous-discrete), this mode of regularization leads us to look for:

$$\widehat{\widehat{x}}^{(\alpha)} = \underset{\widehat{x}\in L^2_{\mathbb{C}}[0,1]}{\arg\min}\left(\|\boldsymbol{x}_N - \boldsymbol{y}\|^2 + \alpha\int_0^1 |\widehat{x}(\nu)|^2\,d\nu\right)\,,$$

where $\boldsymbol{x}_N = [x_1,\ldots,x_N]^T$ and $x_k = \int_0^1 \widehat{x}(\nu)\,\exp\{2j\pi k\nu\}\,d\nu$. The solution reads:

$$\widehat{\widehat{x}}^{(\alpha)}(\nu) = \frac{1}{\alpha+1}\,\widehat{\widehat{x}}^{\text{GI}}(\nu)\,.$$

The spectrum thus regularized is therefore proportional to the periodogram (1.7) that is obtained as a limit case ($\alpha \to 0$). Hence, this type of regularization is not suitable in this example.

However, we will see that the *linear-quadratic* framework above (that combines the *linear* nature of the direct model (1.3) and the *quadratic* nature of the functionals $\mathcal{F}$ and $\mathcal{G}$) turns out to be very handy in practice. The minimization of a criterion such as:

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2_{\mathbf{P}} + \alpha\,\|\boldsymbol{x} - \overline{\boldsymbol{x}}\|^2_{\mathbf{Q}}\,, \tag{2.10}$$

where $\overline{\boldsymbol{x}}$ is a *default* solution (it is the solution obtained when $\alpha \to \infty$, i.e., when the weight given to the data tends towards zero), provides an explicit expression of the minimizer:

$$\widehat{\boldsymbol{x}} = (\mathbf{A}^T\mathbf{P}\mathbf{A} + \alpha\mathbf{Q})^{-1}(\mathbf{A}^T\mathbf{P}\boldsymbol{y} - \mathbf{Q}\,\overline{\boldsymbol{x}}) \tag{2.11}$$

which, thanks to the matrix inversion lemma [SCH 17, SCH 18], can be written:

$$\widehat{\boldsymbol{x}} = \overline{\boldsymbol{x}} + \mathbf{Q}^{-1}\,\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\,\mathbf{A}^T + \alpha^{-1}\,\mathbf{P}^{-1})^{-1}\,(\boldsymbol{y} - \mathbf{A}\,\overline{\boldsymbol{x}})\,. \tag{2.12}$$

In this expression, the matrix to be inverted has, in general, different dimensions from that of (2.11).

2.1.2.2. *Roughness measures*

A very simple way of measuring the roughness of an image is to apply an appropriate difference operator and then calculate the Euclidian norm of the result. As the differentiation operation is linear with respect to the original image, the resulting measure of roughness is quadratic:

$$\mathcal{F}(\boldsymbol{x}) = \|\nabla^k(\boldsymbol{x})\|^2 = \|\mathbf{D}_k\boldsymbol{x}\|^2\,. \tag{2.13}$$

The order k of the difference operator ∇^k is habitually 1 or 2. Measure (2.13) is minimum when $\boldsymbol{x}$ is constant ($k=1$), affine ($k=2$), etc.

2.1.2.3. *Non-quadratic penalization*

Another way of preserving the discontinuities in an object, better than the regularization methods using quadratic criteria, is to use *non-quadratic* penalty functions [IDI 99]. This is precisely what was done to process the interferometry data of Chapter 1, section 1.2, by finally choosing criterion (1.8). The principle is to use a function that increases more slowly than a parabola so as to apply smaller penalties to large variations. These functions are of two main types:

– L_2L_1 functions, i.e., continuously differentiable, convex functions that behave quadratically at the origin and are asymptotically linear. A typical example is the branch of a hyperbola;

– L_2L_0 functions, which differ from the previous ones by being asymptotically constant and thus non-convex.

This time, it is no longer possible to obtain an explicit solution but the first functions have the advantage of being convex, so the standard minimization techniques are sure to converge to the global minimum and give some *robustness* to the solution [BOU 93]. The others enable the discontinuities to be effectively detected but at the expense of some instability and high computing costs [GEM 92].

2.1.2.4. *Kullback pseudo-distance*

In many image processing problems, it is essential to preserve the positivity of the pixel intensities. One way of doing this is to consider that the positive object can be identified, after normalization, with a probability distribution, and then to use the distance measures between probability laws. In particular, the Kullback pseudo-distance (or divergence, or information) of a probability π with respect to a probability π_0 (such that π is absolutely continuous with respect to π_0) can be written:

$$\mathcal{K}(\pi_0, \pi) = \int \left(-\log \frac{d\pi}{d\pi_0} \right) d\pi_0.$$

For a *reference object* whose components m_j are positive,

$$\mathcal{F}(\boldsymbol{x}) = \mathcal{K}(\boldsymbol{x}, \boldsymbol{m}) = \sum_{j=1}^{M} x_j \log \frac{x_j}{m_j} \tag{2.14}$$

is often used. Here again, it is not possible to obtain an explicit expression for the solution; it has to be calculated iteratively [LEB 99].

Criterion (2.5) sums up a view of regularization that can be called *deterministic*, since the only probability law used is, at least implicitly though the choice of the functional $\mathcal{G}$, the one for noise. It has led to important theoretical developments, essentially in mathematical physics. However, the questions of choice of the regularizing functional $\mathcal{F}(x)$ [CUL 79] and the adjustment of the regularization coefficient α [THO 91]

are still very open. Section 2.3 presents the principal methods for adjusting the regularization coefficient that exist in this framework. However, whether we want to set this *hyperparameter* by such a *supervised* method or not, we need to be capable of minimizing the regularized criterion (2.5) in practice afterwards, as a function of $\boldsymbol{x}$. This important aspect of inversion is dealt with below.

2.2. Criterion descent methods

Implicitly or explicitly, most inversion methods are based on the minimization of a criterion. According to the properties of the latter, the computing cost of the solution can vary enormously, typically by a factor of a thousand between the minimization of a quadratic criterion by inversion of a linear system and that of a multimodal criterion by a relaxation technique such as *simulated annealing*, everything else being equal. Finally, the choice of the "right" inversion method depends on the computing facilities available. And we still need to know which algorithm to use for a given optimization problem. For instance, in the comparison above, it would be possible, but completely inefficient, to use simulated annealing to minimize a quadratic criterion. This section gives a non-exhaustive overview of optimization problems in the context of inverse problems in signal and image processing, with the associated algorithms. It obviously cannot replace the literature devoted to optimization as a whole, such as [NOC 99] or [BER 95].

2.2.1. *Criterion minimization for inversion*

By *criterion minimization*, we understand: finding the $\widehat{\boldsymbol{x}}$ that minimizes $\mathcal{J}(\boldsymbol{x})$ among the elements of $\mathcal{X}$. In the rest of this section, we consider the case of real vectors[1]: $\mathcal{X} \subset \mathbb{R}^M$. The criterion $\mathcal{J}$ and the set $\mathcal{X}$ may depend on the data, and structural properties (additional terms in the expression for $\mathcal{J}$ expressing "soft" constraints, whereas the specification of $\mathcal{X}$ is likely to impose "hard" constraints), hyperparameters managing the compromise between fidelity to data and regularity.

Thus, in the case of the generalized inverse of section 1.4, $\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{x}\|$ and $\mathcal{X} = \left\{\boldsymbol{x},\ \mathbf{A}^T\mathbf{A}\boldsymbol{x} = \mathbf{A}^T\boldsymbol{y}\right\}$ is the set of minimizers of $\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|$. In the case of the specification of composite criteria dealt with in section 2.1.2,

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{G}(\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) + \alpha\,\mathcal{F}(\boldsymbol{x}), \tag{2.15a}$$

$$\text{with} \quad \mathcal{X} = \mathbb{R}^M \text{ (non-constrained case)} \tag{2.15b}$$

$$\text{or} \quad \mathcal{X} = \mathbb{R}^M_+ \text{ (positivity constraint)} \tag{2.15c}$$

1. The case where $\boldsymbol{x}$ is a function (more precisely, the case of a space $\mathcal{X}$ of infinite dimension) poses mathematical difficulties that come under functional analysis.

Defining $\widehat{\boldsymbol{x}}$ formally as the minimizer of a criterion hides three main levels of difficulty in terms of implementation. In order of increasing complexity we have:

① $\mathcal{J}$ is quadratic: $\mathcal{J} = \boldsymbol{x}^T \mathbf{M}\, \boldsymbol{x} - 2\, \boldsymbol{v}^T \boldsymbol{x} + \text{const.}$ and $\mathcal{X} = \mathbb{R}^M$, or else $\mathcal{X}$ is affine: $\mathcal{X} = \left\{ \boldsymbol{x}_0 + \mathbf{B}\boldsymbol{u},\ \boldsymbol{u} \in \mathbb{R}^P,\ P < M \right\}$;

② $\mathcal{J}$ is a differentiable convex function and $\mathcal{X} = \mathbb{R}^M$ or a convex (closed) subset of $\mathbb{R}^M$;

③ $\mathcal{J}$ has no known properties.

2.2.2. *The quadratic case*

In situation ①, with $\mathcal{X} = \mathbb{R}^M$ assuming $\mathbf{M}$ is symmetric and invertible, $\widehat{\boldsymbol{x}}$ is the solution of the *linear system* $\mathbf{M}\, \widehat{\boldsymbol{x}} = \boldsymbol{v}$ of dimensions $M \times M$, which expresses the fact that the gradient becomes zero, $\nabla \mathcal{J}(\widehat{\boldsymbol{x}}) = \mathbf{0}$. We have already encountered a similar expression in (2.11) and we will meet it again in the Gaussian linear probabilistic framework of Chapter 3.

In the variant constrained to a space $\mathcal{X}$ that is affine, we need only to replace $\boldsymbol{x}$ by its expression in $\boldsymbol{u}$ to get back to the unconstrained minimization of a quadratic criterion, in $\mathbb{R}^P$.

2.2.2.1. *Non-iterative techniques*

A finite number of operations is sufficient to invert any linear system: of the order of M^3 operations (and M^2 memory locations) for an $M \times M$ system. If the *normal matrix* $\mathbf{M} = \{m_{ij}\}$ has a particular structure, the system inversion cost may decrease.

In signal processing, the stationary nature of a signal is expressed by the Toeplitz character of the normal matrix (i.e., $m_{ij} = \mu_{j-i}$). In image processing using a stationary hypothesis, the normal matrix is Toeplitz-block-Toeplitz (i.e., Toeplitz by blocks, the blocks of sub-matrices themselves being Toeplitz). In both these cases, we find inversion algorithms costing of the order of M^2 operations and M memory locations (Levinson algorithm) and even fast algorithms using a fast Fourier transform costing only of the order of $M \log M$ operations. The spectral expression for the "Wiener filter" of Chapter 4 is a special case where "fast" implementation is possible for the case of a *circulant* normal matrix (i.e., $m_{ij} = \mu_{j-i \bmod M}$).

The *sparse* nature of the normal matrix can also be used to good advantage: if only ML coefficients of $\mathbf{M}$ are non-zero, we can hope to decrease the inversion cost in terms of the number of operations and variables to be stored. For example, if $\mathbf{M}$ is a *band* matrix ($m_{ij} = 0$ if $|j - i| \geq \ell < M$: a band matrix is sparse and L is of the same order as ℓ), the inversion cost does not exceed $M\ell^2$ operations and $M\ell$ memory locations. In particular, a normal matrix $\mathbf{M} = \mathbf{A}^T \mathbf{A}$ is band if $\mathbf{A}$ corresponds to filtering by a small finite impulse response.

2.2.2.2. *Iterative techniques*

If the number of unknowns M is very large (e.g., pixels in image restoration, or voxels for 3D objects), the memory cost of non-iterative techniques often becomes prohibitive. It is then preferable to use a *fixed point* method, iteratively engendering a series $\widehat{\boldsymbol{x}}^{(i)}$ with a limit $\widehat{\boldsymbol{x}} = \mathbf{M}^{-1}\boldsymbol{v}$. All the conventional variants verify $\mathcal{J}(\widehat{\boldsymbol{x}}^{(i+1)}) \leq \mathcal{J}(\widehat{\boldsymbol{x}}^{(i)})$. Three families can be distinguished.

2.2.2.2.1. "Column-action" algorithms

A single component differs between $\widehat{\boldsymbol{x}}^{(i)}$ and $\widehat{\boldsymbol{x}}^{(i+1)}$. The M components are scanned cyclicly during the iterations. This is the principle of the Gauss-Seidel method, or *coordinate descent* [BER 95, p. 143], also called ICM (*iterative conditional modes*) or ICD (*iterative coordinate descent*) in image restoration [BES 86, BOU 93]. It can be generalized for blocks of components and is all the more interesting and partially parallelizable if $\mathbf{A}$ is sparse.

2.2.2.2.2. "Row-action" algorithms

A single component of $\boldsymbol{y}$ is taken into account to calculate $\widehat{\boldsymbol{x}}^{(i+1)}$ from $\widehat{\boldsymbol{x}}^{(i)}$. The N data are scanned cyclicly during the iterations, which makes this approach inevitable if the data are too numerous to be processed simultaneously. The *algebraic reconstruction techniques* (ART), long-standing references in medical imaging by X-ray tomography, follow this principle to minimize the least squares criterion $\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2$ [GIL 72]. They can be generalized to the penalized criterion $\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2 + \alpha \|\boldsymbol{x}\|^2$ [HER 79], can process blocks of data and are all the more interesting and partially parallelizable if $\mathbf{A}$ is sparse.

2.2.2.2.3. "Global" techniques

At each iteration, all the unknowns are updated according to all the data. The *gradient* algorithms are prototypes of the global approach:

$$\widehat{\boldsymbol{x}}^{(i+1)} = \widehat{\boldsymbol{x}}^{(i)} - \lambda(\widehat{\boldsymbol{x}}^{(i)})\, \nabla\mathcal{J}(\widehat{\boldsymbol{x}}^{(i)}),$$

with $\nabla\mathcal{J}(\widehat{\boldsymbol{x}}^{(i)}) = 2\,\mathbf{M}\,\widehat{\boldsymbol{x}}^{(i)} - 2\,\boldsymbol{v}$. Note that the Landweber method defined by (2.4) is in fact a gradient technique minimizing the non-regularized criterion $\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2$. The *conjugate gradient* or *pseudo-conjugate gradient* algorithms are variants that converge more rapidly, in which the successive descent directions combine the previously calculated gradients to avoid the zigzag trajectory of the simple gradient [PRE 86, p. 303]. These variants are of first order, thus occupying little memory; they use only the M first derivatives $\partial\mathcal{J}/\partial x_m$. The preconditioning technique can further increase the efficiency of CG algorithms, as explained in Chapter 4, section 4.4.4, in the context of deconvolution.

We will end with second order techniques. In situation ①, with $\mathcal{X} = \mathbb{R}^M$, and taking $\mathbf{M}$ to be symmetric and invertible, each iteration of the standard Newton's

method can be written:

$$\widehat{\boldsymbol{x}}^{(i+1)} = \widehat{\boldsymbol{x}}^{(i)} - \left(\nabla^2 \mathcal{J}(\widehat{\boldsymbol{x}}^{(i)})\right)^{-1} \nabla \mathcal{J}(\widehat{\boldsymbol{x}}^{(i)}) = \mathbf{M}^{-1}\boldsymbol{v},$$

considering that $\nabla \mathcal{J}(\boldsymbol{x}) = 2\mathbf{M}\boldsymbol{x} - 2\boldsymbol{v}$ and $\nabla^2 \mathcal{J}(\boldsymbol{x}) = 2\mathbf{M}$. In other words, a single iteration of this algorithm is equivalent to solving the problem itself. Unless $\mathbf{M}$ has a specific structure, the computation cost is prohibitive for most realistic inversion problems. Some *quasi-Newton* variants become iterative again by approaching $\mathbf{M}^{-1}$ by a series of matrices $\mathbf{P}^{(i)}$. The most popular among them is the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method [NOC 99, Chapter 8]. For large-sized problems, the computation burden of such quasi-Newton methods is still too high. A better choice is to resort to *limited-memory* BFGS, which can be seen as an extension of the CG method, in-between first and second order techniques [NOC 99, Chapter 9].

2.2.3. *The convex case*

The quadratic criteria are part of a larger family of functions $\mathcal{J}$ that are convex, i.e., such that $\forall \boldsymbol{x}_1,\ \boldsymbol{x}_2 \in \Omega,\ \theta \in\ (0,1)$,

$$\mathcal{J}(\theta\, \boldsymbol{x}_1 + (1-\theta)\, \boldsymbol{x}_2) \leq \theta\, \mathcal{J}(\boldsymbol{x}_1) + (1-\theta)\, \mathcal{J}(\boldsymbol{x}_2)$$

with $\mathcal{X} = \mathbb{R}^M$. Similarly, $\mathcal{X}$ is a convex set if $\forall \boldsymbol{x}_1,\ \boldsymbol{x}_2 \in \mathcal{X},\ \theta \in\ (0,1)$, we have $\theta\, \boldsymbol{x}_1 + (1-\theta)\, \boldsymbol{x}_2 \in \mathcal{X}$.

The specification that the criteria be convex but not necessarily quadratic gives a wider choice as far as modeling is concerned. The Kullback pseudo-distance (2.14) is convex over $\mathbb{R}_+^M$; the Markov penalty functions $\mathcal{F}(\boldsymbol{x}) = \sum_j \varphi(x_j - x_{j+1})$ are convex over $\mathbb{R}^M$ if φ is a convex scalar function such as $\varphi(x) = \sqrt{\tau^2 + x^2}$, which was used in the spectrometry example of Chapter 1, section 1.2.

The minimization of non-quadratic convex criteria, although more difficult and more costly than the minimization of quadratic criteria, remains altogether compatible with modern computing resources, which explains the increasingly frequent use of convex penalty functions in signal and image restoration [IDI 99]. Let us start by recalling a few fundamental properties of convex criteria [BER 95, App. B]:

– a convex continuous criterion $\mathcal{J}$ is *unimodal*: any local minimum is global and the set of its minimizers is convex;

– if $\mathcal{J}_1,\ \mathcal{J}_2$ are convex and $\alpha_1,\ \alpha_2 \geq 0$, then $\alpha_1 \mathcal{J}_1 + \alpha_2 \mathcal{J}_2$ is convex[2];

2. This property "explains" why we are interested in convexity rather than unimodality: for example, the penalized criterion (2.15a) is convex (so unimodal) if $\mathcal{G}$ and $\mathcal{F}$ are convex, whereas the unimodality of $\mathcal{G}$ and $\mathcal{F}$ would not be enough to guarantee the unimodality of the criterion.

– if $\mathcal{J}$ is *strictly* convex, there exists one and only one minimizer $\widehat{\boldsymbol{x}}$ in any convex $\mathcal{X}$ that is *closed* (i.e., the boundary of $\mathcal{X}$ belongs to $\mathcal{X}$).

On the other hand, if the criterion is non-quadratic, the minimizer $\widehat{\boldsymbol{x}}$ is a function of the data that is generally *neither linear, nor explicit*. Owing to this, the non-iterative inversion techniques for linear systems of section 2.2.2 are no longer valid. In contrast, the three families of iterative techniques based on the successive reduction of the criterion give algorithms that converge towards $\widehat{\boldsymbol{x}}$ if $\mathcal{J}$ is convex and differentiable and if $\mathcal{X} = \mathbb{R}^M$. The case of a criterion that is convex but not differentiable is slightly trickier; modern techniques, known as *interior point techniques*, approach the solution by minimizing a succession of differentiable convex approximations [BER 95, p. 312].

There are also other possible families of convergent techniques: *reweighted least squares*, also called *semi-quadratic* algorithms (see Chapter 6), or the approaches based on maximizing a *dual* criterion [BER 95, HEI 00, LUE 69].

If $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^M$, some adaptation is necessary: *projected gradient* or *conditional gradient* versions in the family of "global techniques" [BER 95, Chapter 2], and techniques of *projection on convex sets* [SEZ 82, YOU 82] in the family of "row-action" techniques. As for "column-action" techniques, they remain particularly simple if the constraints are *separable*, i.e., if $\mathcal{X}$ is a Cartesian product, e.g. the positivity corresponds to $\mathcal{X} = \mathbb{R}_+ \times \cdots \times \mathbb{R}_+$. Finally, certain constrained problems are equivalent to a non-constrained problem in the dual domain, which justifies the use of dual methods.

2.2.4. *General case*

In the case of non-convex criteria, the possible existence of local minima makes the use of descent techniques risky, in the sense that any local minimizer is a possible fixed point for most of these techniques. Whether we have convergence towards $\widehat{\boldsymbol{x}}$ rather than towards a local solution then depends on the initialization. Several strategies can be envisaged for avoiding these local solutions. Apart from exceptional cases, they are notoriously more costly than the descent methods and yet still do not guarantee convergence towards the global minimizer. Without guaranteeing convergence mathematically, some techniques are nevertheless sufficiently robust to avoid aberrant solutions. They then give results that could not have been obtained by minimization of a convex criterion, for applications such as automatic image segmentation or object detection.

Two types of approach can be distinguished. On the one hand we have deterministic methods that, in the absence of mathematical convergence properties, favor robustness. For instance, the principle of *gradual non-convexity* (GNC) [BLA 87, NIK 98]

consists of gradually minimizing a series of criteria using a conventional descent technique, starting with a convex criterion and finishing with the non-convex criterion. The robustness of this technique comes from the quality of the initial solution. Its implementation cost and complexity are relatively low. On the other hand, we have the pseudo-random methods (simulated annealing [GEM 84] and adaptive random search [PRO 84]), which make use of the generation of a large number of random samples to avoid the traps. Simulated annealing has (probabilistic) convergence properties but the high computing cost of such techniques explains why their use is still limited in the signal and image restoration field.

2.3. Choice of regularization coefficient

There are few methods for determining the hyperparameters in the framework of this chapter [THO 91]. The most frequently used are the following.

2.3.1. *Residual error energy control*

One of the most intuitive and oldest ideas for setting the value of α that comes into the regularized, or penalized, criterion (2.5) is to consider α as a Lagrange multiplier in the equivalent problem:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x}) \;\; \text{s.t.} \;\; \mathcal{G}(\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) = c\,. \tag{2.16}$$

The degree of regularization is fixed by the value of c, which can be considered as a statistic for which the probability distribution can be deduced from $p(\boldsymbol{y} \,|\, \boldsymbol{x})$. When $\mathcal{G} = \|\cdot\|^2$ and $\boldsymbol{x}_0$ is the *true* solution, the vector of the *residuals* $\boldsymbol{y} - \mathbf{A}\boldsymbol{x}_0$ follows the law of the noise, which is implicitly taken to be homogeneous, zero-mean, white and Gaussian. It results from this that c/σ^2 is a variable of χ^2 with N degrees of freedom if σ^2 is the variance of the noise. It is then recommended to set c to its expectation value, i.e., $N\sigma^2$. However, such a choice often leads to overregularization of the solution. One explanation is that the regularized solution $\widehat{\boldsymbol{x}}$ inevitably differs from the true solution and that the residual errors $\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}}$ that are effectively calculable to obtain the value of $\mathcal{G}$ do not follow any known distribution. Moreover, in many problems, the graph of the function $\mathcal{G}(\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}}) = \mathcal{G}(\alpha)$ is practically horizontal over a large range of values of α: any error in the estimation of σ^2 thus leads to large variations in the value of α that satisfies constraint (2.16).

2.3.2. *"L-curve" method*

It is also possible to use an alternative method that has proved its worth in linear inverse problems of form (2.5) and in the case where the regularization functional $\mathcal{F}(\boldsymbol{x})$

is quadratic. This is the "L-curve" method [HAN 92]. It consists of using a log-log scale to plot the regularization functional $\mathcal{F}(\widehat{\boldsymbol{x}}(\alpha))$ against the least squares criterion $\|\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}}(\alpha)\|^2$ by varying the regularization coefficient α. This curve generally has a characteristic L shape (whence its name) and the value of α corresponding to the corner of the L provides a good compromise between the contradictory requirements of fidelity to the data and fidelity to the prior information.

To understand why this is so, we know that, if $\boldsymbol{x}_0$ is the exact solution, then the error $\widehat{\boldsymbol{x}}(\alpha) - \boldsymbol{x}_0$ can be divided into two parts: a *perturbation* error due to the presence of the measuring error $\boldsymbol{b}$ and a *regularization* error due to the use of a regularizing operator instead of an inverse operator (see (2.2)). The vertical part of the L-curve, described for low values of α, corresponds to solutions for which $\mathcal{F}(\widehat{\boldsymbol{x}}(\alpha))$ is very sensitive to variations in α, as the measurement error $\boldsymbol{b}$ dominates $\widehat{\boldsymbol{x}}(\alpha)$ and does not satisfy the discrete Picard condition [HAN 92]. The horizontal part of the curve, described for high values of α, corresponds to solutions for which it is the sum of the squares of the residuals $\|\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}}(\alpha)\|^2$ that is the most sensitive to variations of α, since $\widehat{\boldsymbol{x}}(\alpha)$ is dominated by the regularization error as long as $\boldsymbol{y} - \boldsymbol{b}$ satisfies the discrete Picard condition.

2.3.3. *Cross-validation*

In the case where the hyperparameters of problem (2.5) are limited simply to the regularization coefficient and where $\mathcal{F}$ and $\mathcal{G}$ are quadratic, *cross-validation* methods also provide acceptable solutions [GOL 79, WAH 77].

The aim is to find a value of the regularization coefficient α such that the regularized solution:

$$\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \big(\mathcal{G}(\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) + \alpha\, \mathcal{F}(\boldsymbol{x}) \big) \tag{2.17}$$

is as close as possible to the actual object $\boldsymbol{x}$. Let Δ_x be a measure of the distance between $\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})$ and $\boldsymbol{x}$. With the choice of quadratic distances for $\mathcal{F}$ and $\mathcal{G}$, it is natural to also choose a quadratic distance for Δ_x:

$$\Delta_x(\alpha, \boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})\|^2 . \tag{2.18}$$

Δ_x can be interpreted as a loss function measuring the *risk* involved in using $\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})$ instead of $\boldsymbol{x}$. A reasonable method for choosing α would be to choose the value that minimizes this risk on the average, i.e., the mean square error (MSE):

$$MSE(\alpha, \boldsymbol{x}) = \int \Delta_x(\alpha, \boldsymbol{x}, \boldsymbol{y})\, p(\boldsymbol{y} \mid \boldsymbol{x})\, d\boldsymbol{y} \tag{2.19}$$

which is an expectation with respect to the noise probability distribution (2.9). Unfortunately, the solution to this problem:

$$\alpha^{\text{MSE}}(\boldsymbol{y}, \boldsymbol{x}) = \arg\min_{\alpha} MSE(\alpha, \boldsymbol{x}) \tag{2.20}$$

depends on the real object which, obviously, is unknown. As the regularized solution $\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})$ can also be seen as a predictor of the observations through $\widehat{\boldsymbol{y}}(\alpha, \boldsymbol{y}) = \mathbf{A}\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})$, it is possible to measure the difference between the real and predicted observations with the following loss function:

$$\Delta_y(\alpha, \boldsymbol{x}, \boldsymbol{y}) = \|\mathbf{A}\boldsymbol{x} - \mathbf{A}\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})\|^2 . \tag{2.21}$$

The value of α could be obtained by minimizing the corresponding mean risk, which, in this case, is the MSE on the prediction:

$$MSEP(\alpha, \boldsymbol{x}) = \int \Delta_y(\alpha, \boldsymbol{x}, \boldsymbol{y})\, p(\boldsymbol{y} \mid \boldsymbol{x})\, d\boldsymbol{y} \tag{2.22}$$

but, there again, the solution depends on the real object. The difficulty can, however, be overcome because the criterion $MSEP(\alpha, \boldsymbol{x})$ can be estimated by *generalized cross-validation* (GCV). Its basic principle is the following. Let $\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y}^{[-k]})$ be the minimizer of the criterion:

$$\mathcal{J}^{[-k]}(\boldsymbol{x}) = \sum_{n \neq k} |y_n - (\mathbf{A}\boldsymbol{x})_n|^2 + \alpha \, \|\boldsymbol{x}\|^2_{\mathbf{Q}} , \tag{2.23}$$

i.e., the object restored by using all the data *except* sample y_k. It is possible to use $\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y}^{[-k]})$ next to predict the missing data item:

$$\widehat{y}_k^{[-k]}(\alpha) = \left[\mathbf{A}\, \widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y}^{[-k]})\right]_k . \tag{2.24}$$

The method consists of looking for the value of α that minimizes a weighted energy of the prediction error $\alpha^{\text{GCV}} = \arg\min_\alpha V(\alpha)$, with:

$$V(\alpha) = \frac{1}{N} \sum_{k=1}^{N} w_k^2(\alpha) \left(y_k - \widehat{y}_k^{[-k]}(\alpha)\right)^2, \tag{2.25}$$

where the coefficients $w_k^2(\alpha)$ are introduced to avoid criterion (2.25) having undesirable properties, such as a lack of invariance during arbitrary rotations of the observation space, or absence of a minimum. They are given by:

$$w_k(\alpha) = \frac{1 - B_{kk}(\alpha)}{1 - \operatorname{trace}\left(\mathbf{B}(\alpha)\right)/M} ,$$

where B_{kk} is the kth diagonal element of the matrix $\mathbf{B}(\alpha) = \mathbf{A}(\mathbf{A}\mathbf{A}^T + \alpha\mathbf{Q})^{-1}\mathbf{A}^T$. The calculation of the minimum relies on the "linear-quadratic" nature of the problem, which allows a simpler relation to be established:

$$V(\alpha) = \frac{N \, \|(\mathbf{I} - \mathbf{B}(\alpha))\, \boldsymbol{y}\|^2}{\left(\operatorname{trace}\left(\mathbf{I} - \mathbf{B}(\alpha)\right)\right)^2} . \tag{2.26}$$

This clearly shows that the GCV function $V(\alpha)$ is, in fact, the sum of the squares of the residual errors *weighted by a coefficient that depends on* α. This method has interesting asymptotic statistical properties. For example [LI 86], $\widehat{\boldsymbol{x}}(\alpha^{\text{GCV}}, \boldsymbol{y})$ gives almost surely the minimum of $\|\mathbf{A}\boldsymbol{x} - \mathbf{A}\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})\|^2$ when $N \to \infty$. Nevertheless, it has to be understood that such a result is of interest only in the case of parsimonious parameterization of the object sought, with a number M of parameters much smaller than the number N of data points. These asymptotic properties and numerous practical results explain why this method has so often been used in 1-D problems. Its use in image processing is more recent [FOR 93, REE 90].

These methods for choosing the regularization coefficient are only clearly justified in the framework of quadratic regularized criteria. The stochastic extension of Chapter 3 will allow us to go beyond this framework.

2.4. Bibliography

[AND 77] ANDREWS H. C., HUNT B. R., *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[BER 95] BERTSEKAS D. P., *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[BES 86] BESAG J. E., "On the statistical analysis of dirty pictures (with discussion)", *J. R. Statist. Soc. B*, vol. 48, num. 3, p. 259-302, 1986.

[BIA 59] BIALY H., "Iterative Behandlung linearen Funktionalgleichungen", *Arch. Ration. Mech. Anal.*, vol. 4, p. 166-176, 1959.

[BLA 87] BLAKE A., ZISSERMAN A., *Visual Reconstruction*, The MIT Press, Cambridge, MA, 1987.

[BOU 93] BOUMAN C. A., SAUER K. D., "A generalized Gaussian image model for edge-preserving MAP estimation", *IEEE Trans. Image Processing*, vol. 2, num. 3, p. 296-310, July 1993.

[BUR 31] BURGER H. S., VAN CITTERT P. H., "Wahre und Scheinbare Intensitätsventeilung in Spektrallinier", *Z. Phys.*, vol. 79, p. 722, 1931.

[CUL 79] CULLUM J., "The effective choice of the smoothing norm in regularization", *Math. Comp.*, vol. 33, p. 149-170, 1979.

[DIA 70] DIAZ J. B., METCALF F. T., "On iteration procedures for equation of the first kind, $Ax = y$, and Picard's criterion for the existence of a solution", *Math. Comp.*, vol. 24, p. 923-935, 1970.

[FOR 93] FORTIER N., DEMOMENT G., GOUSSARD Y., "GCV and ML methods of determining parameters in image restoration by regularization: fast computation in the spatial domain and experimental comparison", *J. Visual Comm. Image Repres.*, vol. 4, num. 2, p. 157-170, June 1993.

[GEM 84] GEMAN S., GEMAN D., "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, num. 6, p. 721-741, Nov. 1984.

[GEM 92] GEMAN D., REYNOLDS G., "Constrained restoration and the recovery of discontinuities", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, num. 3, p. 367-383, Mar. 1992.

[GIL 72] GILBERT P., "Iterative methods for the three-dimensional reconstruction of an object from projections", *J. Theor. Biol.*, vol. 36, p. 105-117, 1972.

[GOL 79] GOLUB G. H., HEATH M., WAHBA G., "Generalized cross-validation as a method for choosing a good ridge parameter", *Technometrics*, vol. 21, num. 2, p. 215-223, May 1979.

[HAN 92] HANSEN P., "Analysis of discrete ill-posed problems by means of the L-curve", *SIAM Rev.*, vol. 34, p. 561-580, 1992.

[HEI 00] HEINRICH C., DEMOMENT G., "Minimization of strictly convex functions: an improved optimality test based on Fenchel duality", *Inverse Problems*, vol. 16, p. 795-810, 2000.

[HER 79] HERMAN G. T., HURWITZ H., LENT A., LUNG H. P., "On the Bayesian approach to image reconstruction", *Inform. Contr.*, vol. 42, p. 60-71, 1979.

[IDI 99] IDIER J., "Regularization tools and models for image and signal reconstruction", in *3nd Intern. Conf. Inverse Problems in Engng.*, Port Ludlow, WA, p. 23-29, June 1999.

[KAL 03] KALIFA J., MALLAT S., ROUGÉ B., "Deconvolution by thresholding in mirror wavelet bases", *IEEE Trans. Image Processing*, vol. 12, num. 4, p. 446-457, Apr. 2003.

[LAN 51] LANDWEBER L., "An iteration formula for Fredholm integral equations of the first kind", *Amer. J. Math.*, vol. 73, p. 615-624, 1951.

[LEB 99] LE BESNERAIS G., BERCHER J.-F., DEMOMENT G., "A new look at entropy for solving linear inverse problems", *IEEE Trans. Inf. Theory*, vol. 45, num. 5, p. 1565-1578, July 1999.

[LI 86] LI K. C., "Asymptotic optimality of C_{L} and GCV in ridge regression with application to spline smoothing", *Ann. Statist.*, vol. 14, p. 1101-1112, 1986.

[LUC 74] LUCY L. B., "An iterative technique for the rectification of observed distributions", *Astron. J.*, vol. 79, num. 6, p. 745-754, 1974.

[LUC 94] LUCY L. B., "Optimum strategies for inverse problems in statistical astronomy", *Astron. Astrophys.*, vol. 289, num. 3, p. 983-994, 1994.

[LUE 69] LUENBERGER D. G., *Optimization by Vector Space Methods*, John Wiley, New York, NY, 1st edition, 1969.

[NAS 76] NASHED M. Z., *Generalized Inverses and Applications*, Academic Press, New York, 1976.

[NAS 81] NASHED M. Z., "Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory", *IEEE Trans. Ant. Propag.*, vol. 29, p. 220-231, 1981.

[NIK 98] NIKOLOVA M., IDIER J., MOHAMMAD-DJAFARI A., "Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF", *IEEE Trans. Image Processing*, vol. 7, num. 4, p. 571-585, Apr. 1998.

[NOC 99] NOCEDAL J., WRIGHT S. J., *Numerical Optimization*, Springer Texts in Operations Research, Springer-Verlag, New York, NY, 1999.

[PRE 86] PRESS W. H., FLANNERY B. P., TEUKOLSKY S. A., VETTERLING W. T., *Numerical Recipes, the Art of Scientific Computing*, Cambridge University Press, Cambridge, MA, 1986.

[PRO 84] PRONZATO L., WALTER E., VENOT A., LEBRUCHEC J.-F., "A general-purpose global optimizer: implementation and applications", *Mathematics and Computers in Simulation*, vol. 26, p. 412-422, 1984.

[REE 90] REEVES S. J., MERSEREAU R. M., "Optimal estimation of the regularization parameter and stabilizing functional for regularized image restoration", *Opt. Engng.*, vol. 29, p. 446-454, 1990.

[SCH 17] SCHUR I., "Uber Potenzreihen, die im Innern des Einheitskreises beschränkt sind", *J. Reine Angew. Math.*, vol. 147, p. 205-232, 1917.

[SCH 18] SCHUR I., "Uber Potenzreihen, die im Innern des Einheitskreises beschränkt sind", *J. Reine Angew. Math.*, vol. 148, p. 122-145, 1918.

[SEZ 82] SEZAN M. I., STARK H., "Image restoration by the method of convex projections: Part 2 – Applications and numerical results", *IEEE Trans. Medical Imaging*, vol. MI-1, num. 2, p. 95-101, Oct. 1982.

[STA 02] STARK J.-L., PANTIN E., MURTAGH F., "Deconvolution in astronomy: a review", *Publ. Astr. Soc. Pac.*, vol. 114, p. 1051-1069, 2002.

[THO 91] THOMPSON A., BROWN J. C., KAY J. W., TITTERINGTON D. M., "A study of methods of choosing the smoothing parameter in image restoration by regularization", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-13, num. 4, p. 326-339, Apr. 1991.

[TIK 63] TIKHONOV A., "Regularization of incorrectly posed problems", *Soviet. Math. Dokl.*, vol. 4, p. 1624-1627, 1963.

[TIK 77] TIKHONOV A., ARSENIN V., *Solutions of Ill-Posed Problems*, Winston, Washington, DC, 1977.

[TIT 85] TITTERINGTON D. M., "Common structure of smoothing techniques in statistics", *Int. Statist. Rev.*, vol. 53, num. 2, p. 141-170, 1985.

[WAH 77] WAHBA G., "Practical approximate solutions to linear operator equations when the data are noisy", *SIAM J. Num. Anal.*, vol. 14, p. 651-667, 1977.

[YOU 82] YOULA D. C., WEBB H., "Image restoration by the method of convex projection: part 1 – Theory", *IEEE Trans. Medical Imaging*, vol. MI-1, num. 2, p. 81-94, Oct. 1982.

Chapter 3

Inversion within the Probabilistic Framework

There are at least two reasons that encourage us to consider solving inverse problems in a Bayesian framework [DEM 89]. It was in this framework that local energy functions and Markov modeling, which have had a lasting influence on low-level image processing, were introduced. It is also this same framework that provides the most consistent and complete answers to problems left in abeyance in other approaches, such as the choice of hyperparameters or the optimization of a multimodal criterion.

3.1. Inversion and inference

To make the link between inversion and statistical inference more explicit, it is useful at this stage to sum up the analysis carried out in Chapter 1. After discretization, the direct problem takes the general form $A(\boldsymbol{x}, \boldsymbol{y}) = 0$, where A is an operator linking the unknown object $\boldsymbol{x} \in \mathbb{R}^M$ to the experimental data $\boldsymbol{y} \in \mathbb{R}^N$. Often, it even takes the explicit form $\boldsymbol{y} = A(\boldsymbol{x})$ or the linear form $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, $\mathbf{A}$ being a matrix. Inversion, i.e., the calculation of $\boldsymbol{x}$ when $\mathbf{A}$ and $\boldsymbol{y}$ are known, is very often an ill-posed problem in two senses.

Firstly, the operator $\mathbf{A}$ is often singular, in the sense that there is a class $\mathcal{K}$ of solutions $\boldsymbol{x} \in \mathcal{K}$ such that $\mathbf{A}\boldsymbol{x} = 0$ (the kernel $\mathrm{Ker}\,\mathbf{A} = \mathcal{K}$ is thus not empty). Any element of $\mathcal{K}$ can be added to any solution to give another solution and we cannot, therefore, invert the the direct relation to determine $\boldsymbol{x}$ uniquely from $\boldsymbol{y}$. This lack of uniqueness makes the discrete inverse problem ill-posed in Hadamard's sense. This situation occurs whenever the instrument response destroys part of the information

Chapter written by Guy DEMOMENT and Yves GOUSSARD.

necessary for the reconstruction of the object. Let us not forget, however, that this ambiguity can be removed by using a more or less empirical rule for choosing among all the solutions, such as taking the minimum norm solution, for example.

Secondly, and more critically, no experimental device is completely free of uncertainty, the simplest source being the finite accuracy of the measurements. It is thus more realistic to consider that the object sought and the measurements taken are connected by an equation of the form $\boldsymbol{y} = A(\boldsymbol{x}) \diamond \boldsymbol{b}$, in which A is an operator describing the essential part of the experiment and $\diamond\, \boldsymbol{b}$ accounts for the the deterioration of this ideal representation by various sources of error (of discretization, measurement, etc.) grouped together in the *noise* term. When the observation mechanism can be approximated by a linear distortion and the addition of noise, this equation reduces to (1.3): $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{b}$. The presence of this noise has the effect of "spreading" the set $\mathcal{K}$, since any element $\boldsymbol{x}$ such that $\mathbf{A}\boldsymbol{x} = \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is "small" relative to the assumed level of noise, can be added to any possible solution to obtain another acceptable solution. However, above all, if the ambiguity is removed by taking a rule for choosing an acceptable solution, it is observed in practice that the latter behaves in an unstable way; small changes in the data entail large variations in the calculated solution. This can easily happen even when the solution is unique and depends continuously on the data, i.e., when the problem is well-posed in Hadamard's sense. In fact, the instability comes from the fact that $\mathbf{A}$ is *ill-conditioned* (see section 1.5).

So we see that, in ill-posed problems, obtaining a solution is not so much a problem of mathematical deduction as a problem of *inference*, i.e., of information processing, which can be summed up in the following question: "how can we draw the best possible conclusions from the incomplete information at our disposal?"

To be acceptable, any scientific inference method should: 1) take *all the available pertinent information* into account; 2) carefully *avoid* assuming information is available when it is not. Probabilistic modeling is a handy, consistent way of describing a situation of incomplete information. We will now see how it leads to a Bayesian statistical approach.

3.2. Statistical inference

It should be made clear from the start that any problem dealt with through a Bayesian approach has to be *well-posed* in the sense that enough information must be provided to allow the probability distributions needed for the calculation to be attributed without ambiguity. This means, at least, that an exhaustive set of possibilities must be specified at the start of each problem. We will call this the *data space* (or *proof space*) if it concerns possible results of the experiment, or the *hypothesis space* if it specifies the hypotheses that we wish to verify. It is also useful to distinguish between two classes of problems, called *estimation* and *choice of model*. The first

studies the consequences of choosing a particular model that is assumed "true", while the aim of the choice of model is to select one model by comparison with one or more other possible candidates.

In an estimation problem, we assume that the model is true for *one* (unknown) value $\boldsymbol{x}_0$ of its parameters and we explore the constraints imposed on the parameters by the data. The hypothesis space is thus the set of all possible values of the parameters $\mathcal{H} = \{x_i\}$. The data consist of one or more samples. For the problem to be well-posed, the space of all the possible samples, $\mathcal{S} = \{z_i\}$, must also be stated. The spaces $\mathcal{H}$ and $\mathcal{S}$ can both be discrete or continuous.

Before making the estimation, it is necessary to state a *logical environment* I which defines our working framework (hypothesis space, data space, relationships between parameters and data, any additional information). Typically, I is defined as a logical proposition stating:

– that the true value of the parameter is in $\mathcal{H}$;

– that the observed data consist of N samples of the space $\mathcal{S}^N$;

– how the parameters are connected with the data (this is the role of the direct model A);

– any additional information.

Of course, the physical nature of the parameters and data is implicitly specified in $\mathcal{H}$, $\mathcal{S}$ and A. Implicitly, all the developments that follow will be within the framework defined by I, which signifies that any probability distribution will be conditioned by I. This conditioning will not be indicated explicitly in order to lighten the notation.

We can now get started on the estimation problem by calculating the probability that each of the possible values of the parameter is the actual value. Let D designate the proposition affirming the values of the experimental data actually observed and H the proposition $\boldsymbol{x}_0 = \boldsymbol{x}$ affirming that one of the possible values of the parameter $\boldsymbol{x}$ is the actual value $\boldsymbol{x}_0$.

3.2.1. *Noise law and direct distribution for data*

In any statistical inference method intended to solve a problem such as (1.3), it is necessary to start by choosing a probability law $q(\boldsymbol{b})$ describing our information – or our uncertainty – on the errors $\boldsymbol{b}$. This is an essential step as it allows the *direct*, or *sampling*, distribution to be found:

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = q\left(\boldsymbol{y} - A(\boldsymbol{x})\right). \tag{3.1}$$

In the vast majority of cases, a centered Gaussian distribution, independent of $\boldsymbol{x}$, is chosen for the errors, which gives:

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) = (2\pi)^{-N/2} \, |\mathbf{R}|^{-1/2} \, \exp\left\{ -\frac{1}{2} \, \|\boldsymbol{y} - A(\boldsymbol{x})\|^2_{\mathbf{R}^{-1}} \right\} ,$$

where $\mathbf{R}$ designates the covariance matrix of the distribution $q(\boldsymbol{b})$. It is often diagonal, or even proportional to the identity. A question arises immediately: What sense is to be given to such a choice and in what situations is such a model appropriate?

With a *frequentist'*s interpretation of a probability, the distribution for the noise should be that of the frequencies of its values in a very large number of repeated measurements. It is then justified by reference to the central limit theorem which says, under fairly broad conditions, that if the noise in a sample of data is the result of a large number of accumulated elementary effects that are "random" and independent, the Gaussian distribution is a good approximation of the real frequency distribution. However, except for fluctuations of electronic origin in a measurement system, the noise is not generally the result of independent effects (think, for example, of the discretization errors that depend on the solution $\boldsymbol{x}_0$). Moreover, to be able to make an inference with this interpretation, it would be necessary for us to have numerous results of other measurements so as to be able to determine these frequencies, which is an extremely rare experimental situation.

This Gaussian "hypothesis" is thus not a hypothesis on the "random" nature of the noise. We are not at all claiming that whatever gives rise to the noise is really random and follows a Gaussian distribution. It is not even a hypothesis in the true sense of the word; it is rather the least compromising – or the most conservative – choice that we can make for the noise distribution in a situation of uncertainty. We are assuming two things here: 1) that the noise can take any real value but that its average value is zero; in other words, there is no systematic measurement error (or if there is, we have been able to detect and correct it), and 2) that we expect there to be a "typical scale" of noise; in other words, large contributions to the noise are not as probable as small ones. To put it another way, we think that the distribution for the noise should have a mean value of zero and a finite standard deviation, even if we have no precise idea of the value of the latter. On the other hand, we have no idea as to the existence or otherwise of cumulants of order greater than two. In these conditions, the least compromising choice with respect to the characteristics that we do not know – which can be justified by information principles [JAY 82] – is that of a Gaussian distribution. In addition, if we suspect that the noise components affecting the N samples have different scales and are correlated, the covariance matrix of the distribution is there to express this hypothesis. It is not necessary to specify its value but if it is unknown, its elements, grouped together in a vector of *hyperparameters* $\boldsymbol{\theta}$, will in general only complicate the problem. They are called *nuisance* parameters for this reason.

This choice is appropriate whenever this information is all we know about the noise. As this is a frequent situation, the choice is often made. If we have additional information about the noise, which leads us to choose a non-Gaussian distribution, we can include it in the same way but the result will be significantly better only if the distribution is very different from a Gaussian one. There are situations – such as imaging with a low particle count – where the data are integers and have low values. Choosing a binomial or Poisson distribution can then improve the results.

3.2.2. *Maximum likelihood estimation*

With simply this direct distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$, we could define the solution of the inverse problem as being that of *maximum likelihood* (ML), the likelihood being the direct distribution in which $\boldsymbol{y}$ takes its observed value and parameter $\boldsymbol{x}$ becomes the variable:

$$\widehat{\boldsymbol{x}}^{\text{ML}} = \underset{\boldsymbol{x} \in \mathcal{H}}{\arg\max}\ p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}).$$

In general, the justification for this choice comes from the "good" statistical characteristics (more often than not asymptotic) of this estimator. The *least squares* solution is the special case of the maximum likelihood solution when the direct distribution is Gaussian:

$$\widehat{\boldsymbol{x}}^{\text{LS}} = \underset{\boldsymbol{x} \in \mathcal{H}}{\arg\min}\ (\boldsymbol{y} - A(\boldsymbol{x}))^T\, \mathbf{R}^{-1}\, (\boldsymbol{y} - A(\boldsymbol{x}))\,.$$

Introduced in this way, it is still a *weighted* least squares method (weighted by the matrix $\mathbf{R}^{-1}$) that possesses the indispensable property of invariance under changes of units in $\mathcal{H}$ and $\mathcal{S}$. In many simple situations, this inference method provides all the information we are looking for. However, in inverse problems where the parameterization is not parsimonious, the direct distribution does not contain all the information needed to make the problem well-posed and it does not provide all the technical apparatus necessary for the calculation:

1) In the special case of an indeterminate linear problem $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, where $\mathbf{A}$ is singular (a problem known as *generalized inversion*), there is no "noise" and so no direct distribution, except in the rudimentary sense where $p(\boldsymbol{y} \mid \boldsymbol{x})$ is constant if $\boldsymbol{x}$ is in the class $\mathcal{C}$ of possible antecedents of $\boldsymbol{y}$, and zero otherwise. As the likelihood is constant in class $\mathcal{C}$, maximizing it is of no help for the choice within this class. The essence of the problem does not lie in the presence of "random" noise perturbing the data, but rather in the fact that our information is incomplete, although essentially noise free.

2) In the linear case (1.3), matrix $\mathbf{A}$ of the direct problem is often *ill-conditioned*. The solving operator $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\,\mathbf{R}^{-1}$ is unstable and the solution $\widehat{\boldsymbol{x}}^{\text{ML}} = \mathbf{A}^\dagger\,\boldsymbol{y}$ is unacceptable: the amplification of the noise is excessive.

3) The problem can have nuisance parameters that are of no interest to us, and they may be numerous. When matrix $\mathbf{R}$ is full, $N\,(N-1)/2$ hyperparameters are added

which, when they are unknown, generally have to be estimated by ML as parameters of interest $\boldsymbol{x}$, and the global maximum may then no longer be a point but a whole region.

4) We may have highly pertinent information on the solution we are looking for. For example, we may know that it has to be positive, or satisfy certain constraints (as in astronomical imaging where the integral of the object may already be known), or that it is made up of homogeneous regions separated by clear boundaries. Such information is not contained in the direct distribution but it would be most unreasonable to ignore it.

5) In many problems, it is necessary to obtain not only a solution but also an indication of the confidence we can have in it. If we simply have the direct distribution (3.1), the *confidence intervals* given by the frequency approach only give us information on the *long term* behavior of the solution, i.e., its average behavior over a very large number of repeats of the experiment. However, we only possess the results of a single experiment, which often cannot be reproduced.

6) Finally, the estimation of the parameters of a model that is assumed to be valid is often just one step and we may need to judge the relative merits of various models.

It is therefore necessary to go beyond inference by ML. All the extensions mentioned above are "automatically" provided by the Bayesian approach.

3.3. Bayesian approach to inversion

Bayesian inference is so named because it makes great use of Bayes' rule, which itself is a consequence of a fundamental rule in probability calculation, the *product rule* [COX 61]. Let H be a hypothesis whose truth we want to evaluate and D a set of data connected with this hypothesis. The product rule stipulates that:

$$\Pr(H,\, D) = \Pr(H \mid D)\ \Pr(D) = \Pr(D \mid H)\ \Pr(H)$$

where, for example, $\Pr(H \mid D)$ usually designates the probability that H is true knowing D. From this we draw Bayes' rule:

$$\Pr(H \mid D) = \Pr(H)\ \Pr(D \mid H) / \Pr(D)$$

which is none other than a *learning* rule. It tells us how we should adjust the probability attributed to the truth of a hypothesis when our state of knowledge changes with the acquisition of data. The probability *a posteriori* for H, $\Pr(H \mid D)$, is obtained by multiplying its probability *a priori*, $\Pr(H)$, by the probability of having observed the data D assuming the hypothesis is true, $\Pr(D \mid H)$, and dividing the whole by the probability of having observed the data independently of whether the hypothesis is true or not, $\Pr(D)$. This last term, sometimes called the *global likelihood*, plays the role of a normalization constant.

A large part of statistical inference is based on the use of prior information on the quantities to be estimated, which adds to the information given by the data. Thus, it is not surprising, if we think about the deep nature of the regularization principle set out in Chapter 2, that it shows a close link with Bayesian inference.

In the case of an inverse problem such as (1.3) and assuming that the probability distributions concerned admit a density, the prior information on object $\boldsymbol{x}$ is expressed, in a Bayesian context, in the form of an *a priori* probability density function (pdf) $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Bayes' rule allows us to combine this with the information contained in the data to obtain the *a posteriori* law:

$$p(\boldsymbol{x} \mid \boldsymbol{y}, A, \boldsymbol{\theta}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{x}, A, \boldsymbol{\theta})}{p(\boldsymbol{y} \mid A, \boldsymbol{\theta})} = \frac{p(\boldsymbol{x}, \boldsymbol{y} \mid A, \boldsymbol{\theta})}{p(\boldsymbol{y} \mid A, \boldsymbol{\theta})}\,. \tag{3.2}$$

In this equation, $\boldsymbol{\theta}$ is a vector of *hyperparameters* composed of the parameters of the *a priori* distributions of the errors and the object, and $p(\boldsymbol{y} \mid \boldsymbol{x}, A, \boldsymbol{\theta})$ designates the data law conditioned by the true solution $\boldsymbol{x}$. It is completely determined by the knowledge of the direct model (1.3) and the noise probability law. The last term ensures the normalization of the *a posteriori* law:

$$p(\boldsymbol{y} \mid A, \boldsymbol{\theta}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, A, \boldsymbol{\theta})\, p(\boldsymbol{x} \mid \boldsymbol{\theta})\; d\boldsymbol{x}\,. \tag{3.3}$$

In the Bayesian approach, the knowledge (or uncertainty) about object $\boldsymbol{x}$ after observation of *data* $\boldsymbol{y}$ *only* is wholly described by the probability distribution (3.2). This probability is equal, with just a multiplying factor, to the product of the likelihood introduced in section 3.2 by the *a priori* probability $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. If we assume that, in the case of section 3.2, the knowledge of $\boldsymbol{x}$ (which then comes purely from observations and from the structure of the problem) is represented by the likelihood, we observe that, in the Bayesian approach, taking prior information into consideration by means of $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ modifies our knowledge and, in general, has the effect of reducing the uncertainty on the parameter $\boldsymbol{x}$. But above all, because of the framework adopted, the Bayesian approach enables a wider range of answers to the question "*given a probability distribution for a continuous or discrete parameter* $\boldsymbol{x}$*, what best estimate can be made and with what accuracy?*". There is not a single answer to this question; the problem concerns the theory of the decision that answers the question "*what should we do?*". This implies value judgements and consequently goes beyond the principles of inference, which only answers the question "*what do we know?*". Thus, we can equally well deduce a point estimator or a region of uncertainty from (3.3) [MAR 87, TAR 87]. The *maximum a posteriori* is a frequent choice for the estimator. It consists of giving $\boldsymbol{x}$ the value that maximizes the distribution *a posteriori*:

$$\widehat{\boldsymbol{x}}^{\text{MAP}} = \arg\max p(\boldsymbol{x} \mid \boldsymbol{y}, A, \boldsymbol{\theta})\,. \tag{3.4}$$

However, this is only one of the possible solutions. This MAP estimation corresponds to the minimization of a mean decision cost with an all-or-nothing cost function, the

limit (when $\varepsilon \to 0$) of the mean cost $\Pr(\| \widehat{\boldsymbol{x}} - \boldsymbol{x}_0 \| > \varepsilon)$. Other cost functions have been proposed in the framework of image modeling by Markov fields. They lead to the maximization of the marginal probabilities [BES 86, MAR 87].

3.4. Links with deterministic methods

In the case that interests us here, i.e., an inverse problem in a finite dimension, it is clear that regularizing according to the general principle indicated in Chapter 2, and thus minimizing a criterion such as (2.5), is equivalent to choosing the solution that maximizes the following *a posteriori* law:

$$p(\boldsymbol{x} \,|\, \boldsymbol{y}, A, \boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2\,\sigma^2}\big(\mathcal{G}\,(\boldsymbol{y} - A(\boldsymbol{x})) + \alpha\,\mathcal{F}(\boldsymbol{x})\big)\right\}. \tag{3.5}$$

where σ^2 is the variance of the noise. The above probability law is only one of the possible choices since any strictly monotonic function other than an exponential would do. However, this choice is particularly suitable here because, with the linear model (1.3), taking the usual hypotheses that the noise is Gaussian and independent, as $\mathcal{G}$ is a Euclidian norm, the conditional law $p(\boldsymbol{y} \,|\, \boldsymbol{x}, A, \boldsymbol{\theta})$ is really:

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}, A, \boldsymbol{\theta}) \propto \exp\left\{\frac{1}{2\sigma^2}\mathcal{G}\,(\boldsymbol{y} - A(\boldsymbol{x}))\right\}. \tag{3.6}$$

For the analogy to be complete, the *a priori* law must take the following form:

$$p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \propto \exp\left\{-\frac{\alpha}{2\sigma^2}\mathcal{F}(\boldsymbol{x})\right\}, \tag{3.7}$$

and, for it to be rigorous, the *a posteriori* law (3.5) must be proper, a sufficient condition being that (3.6) and (3.7) are also proper:

$$\int_{\mathbb{R}^N} \exp\left\{-\frac{1}{2\sigma^2}\mathcal{G}\,(\boldsymbol{y} - A(\boldsymbol{x}))\right\} d\boldsymbol{y} < +\infty\,, \quad \int_{\mathbb{R}^M} \exp\left\{-\frac{\alpha}{2\sigma^2}\mathcal{F}(\boldsymbol{x})\right\} d\boldsymbol{x} < +\infty\,.$$

Many local energy functions used in image processing were introduced in a Bayesian framework. They define $\boldsymbol{x}$ as a Markov field (see Chapter 7). Although the *energy* point of view is also held by some members of the image processing community, criteria of form (2.5) can generally be reinterpreted in a Bayesian framework, even if it means making minor changes in $\mathcal{F}$ to ensure the normalization of equation (3.7).

In consequence, the maximum *a posteriori* estimator, which is the Bayesian estimator the most used in inversion, becomes the same as the minimizer of the penalized criterion (2.5):

$$\begin{aligned}\widehat{\boldsymbol{x}}^{\text{MAP}} &= \arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \,|\, \boldsymbol{y}, A, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y} \,|\, A, \boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{x}} \mathcal{G}(\boldsymbol{y} - A(\boldsymbol{x})) + \alpha\,\mathcal{F}(\boldsymbol{x})\end{aligned}$$

under the technical conditions that allow this development (principally, that the problem brings in a finite number of variables). It is thus obvious that the Bayesian framework gives a statistical sense to the minimization of penalized criteria. The question is not, however, whether the Bayesian approach is a justification of the other approaches. We could also, and conversely, say that the same result gives a deterministic interpretation of the probabilistic estimator of the maximum *a posteriori* and that an estimator, once defined, depends no more on the formal framework that engendered it than on the digital means used to calculate it. The question is rather one of seeing that the Bayesian approach provides an answer to the problems raised in section 3.2. In addition to its great consistency, it makes original tools available:

– *marginalization* (everything that does not interest us is simply integrated out of the problem);

– *regression* (the conditional expectation does not have an equivalent in the energy framework);

– *stochastic sampling* (Monte Carlo methods, simulated annealing algorithms, genetic algorithms), not conceivable without the Bayesian approach (on this point, see Chapter 7, section 7.4.2).

3.5. Choice of hyperparameters

The Bayesian framework appreciably extends the range of methods available for determining the hyperparameters. To be applied effectively, all the methods described in Chapter 2 require us to choose the value of the regularization coefficient α and, more generally, all hyperparameters $\boldsymbol{\theta}$ defining the $\mathcal{F}$ and $\mathcal{G}$ distance measures: the variance of the noise, the object correlation parameters and the parameters of the local energy functions. The determination of $\boldsymbol{\theta}$ is the most delicate step in image restoration and reconstruction methods. Although the problem is still open, the Bayesian approach provides consistent tools for tackling it.

Hyperparameters $\boldsymbol{\theta}$ constitute a second level in the description of the problem, which is indispensable to "rigidify" the first level composed of the parameters themselves – i.e., the object $\boldsymbol{x}$. In an ill-posed problem, the value of the parameters is important for obtaining an acceptable solution but has no intrinsic interest. In a Bayesian approach, two levels of inference can be distinguished. The first is inference on $\boldsymbol{x}$, for a given value of $\boldsymbol{\theta}$, through the *a posteriori* distribution of equation (3.2). The second is inference on $\boldsymbol{\theta}$ through the analog relationship:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, A) = p(\boldsymbol{\theta} \mid A)\, p(\boldsymbol{y} \mid \boldsymbol{\theta}, A)/p(\boldsymbol{y} \mid A)\,. \tag{3.8}$$

Here again, we find a characteristic of the use of Bayes' rule: the marginal likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta}, A)$ attached to the data in the second level is the coefficient of normalization in the first.

If, as is often the case, this term is sufficiently peaked, i.e., if the data $\boldsymbol{y}$ contain enough information, the influence of the *a priori* distribution $p(\boldsymbol{\theta} \mid A)$ is negligible and the second level of inference can be solved by maximizing the likelihood. But to do this, we have to solve the marginalization problem corresponding to the calculation of the integral in (3.3). Such integrals rarely lead to an explicit result. One notable exception is the joint Gaussian distribution $p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}, A)$, as we will see in section 3.8.

To get around the problem posed by the explicit calculation of a marginal likelihood, we can introduce "hidden variables" $\boldsymbol{q}$ which complete the observations $\boldsymbol{y}$ in such a way that the new likelihood $p(\boldsymbol{y}, \boldsymbol{q} \mid \boldsymbol{\theta}, A)$ is simpler to calculate. We are then led to maximize the conditional expectations by iterative, deterministic or stochastic techniques (EM and SEM algorithms) [DEM 77], the algorithm converging towards the solution of ML. The need for such stochastic approaches appeared when it was found to be impossible to implement convergent likelihood maximization methods by conventional optimization techniques, as the likelihood was not calculable.

Furthermore, the joint distribution or *generalized likelihood*:

$$p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta}, A) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}, A)\, p(\boldsymbol{y} \mid \boldsymbol{\theta}, A) = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, A)\, p(\boldsymbol{x} \mid \theta) \tag{3.9}$$

sums up all the information at the first level of inference. Its maximization with respect to $\boldsymbol{x}$ and $\boldsymbol{\theta}$ can be envisaged. Thus, the integration problem raised by (3.3) is obviously removed. At fixed $\boldsymbol{\theta}$, the *generalized maximum likelihood* (GML) coincides with the MAP; at fixed $\boldsymbol{x}$, it corresponds to the usual ML for $\boldsymbol{\theta}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ being known. Nevertheless, repeated alternation of these two steps is hazardous: the characteristics of the corresponding estimator are not those of the usual ML [LIT 83]. It can even happen sometimes that the GML is not defined because the likelihood may have no maximum, even local [GAS 92]. This technique thus has a marked empirical character.

Thus, the Bayesian approach leads fairly naturally to the use of estimators based on likelihood for the estimation of the hyperparameters. Despite definite difficulties of implementation, interesting results have been obtained in a one-dimensional framework. In a two- or three-dimensional framework, we have to be more cautious. Although it is possible to estimate the hyperparameters in several cases, the values obtained using this approach do not necessarily lead to good results for the estimation of the parameter of interest $\boldsymbol{x}$, particularly when the latter comes from "natural" data. The cause could lie in there being too great a difference between these natural data and the behavior of the *a priori* model. The question of hyperparameter estimation thus remains wide open.

3.6. *A priori* model

A reproach that is often levelled against Bayesian estimation is that it depends on the knowledge of a hypothetical, uncertain "true model" that engendered the object

to be reconstructed. To formulate this reproach, we have to implicitly accept that reality can be "enclosed" in a mathematical model. This opens up a huge philosophical debate... In the case of the probabilistic approach to inversion as we see it, the frequentist's interpretation of the probabilities maintains an annoying confusion. It does, however, seem important to recall that our probabilistic hypotheses are not hypotheses on the "random" character of the object but choices of a way of representing incomplete prior information – or uncertain knowledge – compatible with the chosen inference tool. This situation is far from unusual, as it is rare for the prior information available in a real problem to come in a form directly suited to the theoretical framework chosen for its processing.

Let us remember that the advantages of the Bayesian approach stem not so much from the additional information introduced by the prior – the energy and deterministic interpretations of the functional of regularization $\mathcal{F}(\boldsymbol{x})$ of section 3.4 show that this information is not proper to the Bayesian approach, and the information on nuisance parameters is *diffuse* most of the time – as from the access it provides to a layer of tools that does not exist in the other approaches, such as marginalization, regression and pseudo-random algorithms.

Having said this, the conversion of prior information into probabilities is a tricky problem that is still far from being solved. To describe object $\boldsymbol{x}$, the prior is often chosen *pragmatically*, as we will see later. There are, however, some *formal rules* that lead to reasonable choices [BER 94, KAS 94, ROB 97] and are used in particular for the hyperparameters. They often lead to an *improper law*, which does not cause any special difficulty if it is handled correctly [JEF 39]. Here are a few examples.

Some methods rely on *transformation group* theory to determine the "natural" reference measure for the problem and to satisfy certain invariance principles. In practice though, this approach has done little more than justify the use of Lebesgue's method for the *localization* parameters (thus providing an extension to the continuous case of the uniform distribution resulting from the application of Bernouilli's "indifference principle" in the discrete case) and the Jeffreys measure in the case of *scale* parameters [JEF 39, POL 92].

Other methods are based on information principles. These are mainly *maximum entropy* methods (MEM), in which we look for the distribution that is closest to the reference distribution (in the Kullback divergence sense) whilst verifying incomplete prior information [JAY 82]. There again, this approach has mainly just helped to justify certain choices after the event. In addition, it is only really workable when the prior information is made up of linear constraints on the distribution we are looking for (moments). We are thus working in the family of *exponential distributions*.

Another formal principle consists of using a *conjugate prior*, i.e., a prior belonging to the same family as the direct distribution of the problem, to obtain an *a posteriori*

distribution in the family [ROB 97]. This is only of interest if the family in question is as small as possible and parametrized. In this case, the step from *a priori* to *a posteriori*, by application of Bayes' rule, comes down to updating the parameters. The interest of this method is essentially technical, as the *a posteriori* is always calculable, at least up to a certain point. A partial justification can also be found by invariance reasoning: if the data $\boldsymbol{y}$ change $p(\boldsymbol{x})$ into $p(\boldsymbol{x} \mid \boldsymbol{y})$, the information that $\boldsymbol{y}$ contributes about $\boldsymbol{x}$ is clearly limited; it should not lead to a change of the whole structure of $p(\boldsymbol{x})$, but only of its parameters. It is obvious though that the main motivation for using the method is its convenience. However, only certain families of direct distributions, such as *exponential families* [BRO 86], guarantee the existence of conjugate priors and it is often necessary to limit use of the method to this class of distributions. In addition, the "automatic" nature of this way of making choices is rather deceptive because additional *hyperparameters* – the values of which have to be specified – inevitably appear.

A last, very important class is composed of "tailor made" constructions, in other words, constructions that are not based on general principles like the previous ones but make pragmatic use of probabilistic methods that express the properties expected of the solutions as well as possible. It is into this category that we must put the Gibbs-Markov fields, which have undergone spectacular development in imaging since 1984 [GEM 84] and which allow essential local properties that an object must possess to be incorporated into an *a priori* distribution. The construction of these models requires considerable know-how but is a very powerful way of incorporating elaborate prior information. The price to be paid for this is high complexity, both in the handling of the models and in the implementation of the resulting estimators. Chapter 7 is entirely devoted to Gibbs-Markov models.

3.7. Choice of criteria

The Bayesian approach brings inversion down to the determination of an *a posteriori* law. Since we cannot envisage calculating such laws completely, we content ourselves with looking for a point estimator, which is often the maximum *a posteriori* one. There are alternatives (*marginal maximum a posteriori*, *mean a posteriori,* etc.) but it is important to assess the consequences of such a choice carefully and, if necessary, think about alternatives.

It is reasonable to raise the question of the necessity for the solution to be continuous with respect to the data and, consequently, the need for convexity of the regularization criteria. While quadratic and entropy approaches are well known for making inverse problems well-posed, the minimization of a non-convex functional cannot guarantee that the solution will be continuous: a small variation in the data can induce a "jump" from one valley to another and thus a loss of continuity. However, in many

problems, these transitions are not only desirable but necessary to restore discontinuities, edges, interfaces, bright spots, etc. without limits in terms of spatial resolution. We can shed a different light on this problem by noting that certain non-convex criteria introduced in imaging have an equivalent expression implying *hidden variables*. In this case, the problem leaves convex analysis and incorporates a measure of combinatory analysis or hypothesis testing, which comes more under decision theory than estimation. Bayesian analysis remains pertinent in this combined detection-estimation context. Much recent work has followed this direction, combining several levels of variables, mixing low- and high-level descriptions, or data acquired by different experimental means. It is in this sense that the conventional concepts of regularization, such as continuity with respect to the data, are not completely appropriate and an effort should be made to extend them.

3.8. The linear, Gaussian case

The Gaussian laws associated with linear direct models provide a linear estimation structure and thus a very convenient algorithmic framework. However, they only allow us to incorporate crude information, basically limited to second order characteristics. Thus, in standard regularization theory [TIT 85], the choice of a quadratic term for fidelity to the data: $\mathcal{G}(\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2_{\mathbf{P}}$ is equivalent to choosing a Gaussian distribution for the noise: $q(\boldsymbol{b} \,|\, \mathbf{R}_b) \sim \mathcal{N}(0, \mathbf{R}_b)$, with $\mathbf{R}_b \propto \mathbf{P}^{-1}$. Similarly, choosing quadratic penalization: $\mathcal{F}(\boldsymbol{x}) = \|\mathbf{D}_k\boldsymbol{x}\|^2$ is also equivalent to choosing a Gaussian prior distribution for the object: $p(\boldsymbol{x} \,|\, \mathbf{R}_x) \sim \mathcal{N}(0, \mathbf{R}_x)$, with $\mathbf{R}_x \propto (\mathbf{D}_k^T\mathbf{D}_k)^{-1}$, assuming, of course, that the matrix $\mathbf{D}_k^T\mathbf{D}_k$ is defined as positive. Deterministic "linear-quadratic" regularization is thus rigorously equivalent to Gaussian linear estimation and the solution, which is explicit, is given by equations (2.11) and (2.12):

$$\widehat{\boldsymbol{x}} = (\mathbf{A}^T\mathbf{R}_b^{-1}\mathbf{A} + \mathbf{R}_x^{-1})^{-1}\mathbf{A}^T\mathbf{R}_b^{-1}\boldsymbol{y}\,, \tag{3.10}$$

$$= \mathbf{R}_x\,\mathbf{A}^T\,(\mathbf{A}\,\mathbf{R}_x\,\mathbf{A}^T + \mathbf{R}_b)^{-1}\,\boldsymbol{y}\,, \tag{3.11}$$

and has the remarkable characteristic of being a linear function of data $\boldsymbol{y}$. This "linear-quadratic" or linear Gaussian inversion holds a dominant position in inversion problems and it is a common reaction to say *"inverse problems aren't complicated; you just need to smooth the data before doing the inversion"*. This way of seeing things is not wrong and is, in fact, sufficient for many problems but it is limiting; it stops us from going further and induces a cascading scheme – linear filtering of a generalized inverse solution – that is only justified in the "linear-quadratic" framework.

3.8.1. *Statistical properties of the solution*

Solution (3.10) is, in the Gaussian case, the mode, the mean and the median of the *a posteriori* probability distribution (3.5) all at once. It minimizes several very

commonly used cost criteria, in particular the *mean quadratic error*. Obviously, in this case, we are talking about a *mean with respect to the a posteriori distribution*, but many physicists and engineers only know the mean square error (MSE) defined as a *mean with respect to the direct distribution* (3.6). It is therefore useful to study the MSE, which is the sum of the bias energy and the trace of the covariance matrix: $MSE(\widehat{\boldsymbol{x}}) = \left\|\mathrm{E}(\widehat{\boldsymbol{x}}) - \boldsymbol{x}_0\right\|^2 + \operatorname{trace}\mathrm{Cov}(\widehat{\boldsymbol{x}})$, designating the "true" solution by $\boldsymbol{x}_0$. For the sake of simplicity, we will assume that the noise is stationary and white: $\mathbf{R}_b = \sigma_b^2\,\mathbf{I}$ and that we can write $\mathbf{R}_x = \sigma_x^2\,(\mathbf{D}^T\mathbf{D})^{-1}$. We thus have $\alpha = \sigma_b^2/\sigma_x^2$.

The expectation of regularized solution (2.11), for direct distribution (3.6), can be written:

$$\begin{aligned}\mathrm{E}(\widehat{\boldsymbol{x}}) &= \mathrm{E}\big((\mathbf{A}^T\mathbf{A} + \alpha\,\mathbf{D}^T\mathbf{D})^{-1}\,\mathbf{A}^T(\mathbf{A}\boldsymbol{x}_0 + \boldsymbol{b})\big)\\ &= (\mathbf{A}^T\mathbf{A} + \alpha\,\mathbf{D}^T\mathbf{D})^{-1}\,\mathbf{A}^T\mathbf{A}\boldsymbol{x}_0\,.\end{aligned}$$

Thus, for the bias to be zero ($\mathrm{E}(\widehat{\boldsymbol{x}}) - \boldsymbol{x}_0 = 0$), we would need $\alpha = 0$, i.e., we must not regularize! The bias energy is:

$$\left\|\mathrm{E}(\widehat{\boldsymbol{x}}) - \boldsymbol{x}_0\right\|^2 = \left\|\left((\mathbf{A}^T\mathbf{A} + \alpha\,\mathbf{D}^T\mathbf{D})^{-1}\,\mathbf{A}^T\mathbf{A} - \mathbf{I}\right)\boldsymbol{x}_0\right\|^2,$$

an increasing function of α, that equals zero and has a zero derivative at $\alpha = 0$ and that tends towards $\left\|\boldsymbol{x}_0\right\|^2$ when $\alpha \to \infty$.

The covariance matrix of the solution can be written:

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{x}}) &= \mathrm{E}\big((\widehat{\boldsymbol{x}} - \mathrm{E}(\widehat{\boldsymbol{x}}))\,(\widehat{\boldsymbol{x}} - \mathrm{E}(\widehat{\boldsymbol{x}}))^T\big)\\ &= \sigma_b^2\,(\mathbf{A}^T\mathbf{A} + \alpha\,\mathbf{D}^T\mathbf{D})^{-1}\,\mathbf{A}^T\mathbf{A}\,(\mathbf{A}^T\mathbf{A} + \alpha\,\mathbf{D}^T\mathbf{D})^{-1}\,.\end{aligned}$$

To calculate its trace, we assume that matrices $\mathbf{A}$ and $\mathbf{D}$ have the same singular vectors[1], so that we have the factorizations:

$$\mathbf{A}^T\mathbf{A} = \mathbf{U}\,\boldsymbol{\Lambda}_a^2\,\mathbf{U}^T \quad \text{and} \quad \mathbf{D}^T\mathbf{D} = \mathbf{U}\,\boldsymbol{\Lambda}_d^2\,\mathbf{U}^T,$$

where $\boldsymbol{\Lambda}_a$ and $\boldsymbol{\Lambda}_d$ are diagonal matrices composed respectively of the singular values $\lambda_a(k)$ of $\mathbf{A}$ and $\lambda_d(k)$ of $\mathbf{D}$, $k = 1,\, 2,\, \ldots,\, M$. We thus obtain:

$$\operatorname{trace}\mathrm{Cov}(\widehat{\boldsymbol{x}}) = \sigma_b^2 \sum_{k=1}^{M} \frac{\lambda_a^2(k)}{\left(\lambda_a^2(k) + \alpha\,\lambda_d^2(k)\right)^2}\,,$$

a strictly decreasing function of α, tending towards zero when $\alpha \to \infty$.

1. This is the case, for example, if $\mathbf{D}$ is the identity matrix, or if $\mathbf{A}$ and $\mathbf{D}$ are two circulant matrices, such as those we will encounter in Chapter 4.

Thus, there is an "optimum", strictly positive, value of α, that makes the MSE minimum. It is worth noting, however, that it depends on the true solution $\boldsymbol{x}_0$ through the bias energy and that efforts to find it would therefore be in vain. Also note that an approach, frequent in statistics, consisting of looking for estimators without bias and, if a degree of freedom remains, with minimum variance[2], leads to the generalized inverse solution, the MSE of which is:

$$MSE(\widehat{\boldsymbol{x}}^{\text{GI}}) = \operatorname{trace} \operatorname{Cov}(\widehat{\boldsymbol{x}}^{\text{GI}}) = \sigma_b^2 \sum_{k=1}^{M} \frac{1}{\lambda_a^2(k)} \,.$$

This can be considerable when some singular values $\lambda_a(k)$ are small, which is precisely the case in discretized and ill-conditioned problems. It can thus be said that, in terms of MSE, regularization consists of voluntarily introducing a bias in order to considerably reduce the variance of the solution.

3.8.2. *Calculation of marginal likelihood*

The linear, Gaussian case is one of the few that allow an explicit calculation of the marginal likelihood of equation (3.3), used to adjust the values of hyperparameters $\boldsymbol{\theta}$. When these are limited to the variances σ_b^2 and σ_x^2 for example (or to the pair σ_b^2 and $\alpha = \sigma_b^2/\sigma_x^2$), we have:

$$p(\boldsymbol{x},\, \boldsymbol{y} \,|\, \sigma_x^2, \sigma_b^2) = \left(2\pi\, \sigma_b^2\right)^{-N/2} \left(2\pi\, \sigma_x^2\right)^{-M/2} \left|\mathbf{D}^T\mathbf{D}\right|^{1/2} e^{-\mathcal{Q}/2\sigma_b^2} \,,$$

$$\text{where} \quad \mathcal{Q} = (\boldsymbol{y} - \mathbf{A}\boldsymbol{x})^T\, (\boldsymbol{y} - \mathbf{A}\boldsymbol{x}) + \alpha\, \boldsymbol{x}^T \mathbf{D}^T \mathbf{D}\, \boldsymbol{x} \,.$$

To calculate the ordinary, or marginal, likelihood of α and σ_b^2, we have to "*integrate $\boldsymbol{x}$ out of the problem*". To prepare this integration, a perfect square is conventionally made to appear in $\mathcal{Q}$:

$$\mathcal{Q} = (\boldsymbol{x} - \widehat{\boldsymbol{x}})^T\, (\mathbf{A}^T\mathbf{A} + \alpha\, \mathbf{D}^T\mathbf{D})\, (\boldsymbol{x} - \widehat{\boldsymbol{x}}) + \mathcal{S}(\alpha) \,,$$

with $\mathcal{S}(\alpha) = \boldsymbol{y}^T(\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}})$, which leads to a Gaussian integral:

$$\begin{aligned} p_{\boldsymbol{y}} \,|\, \alpha,\, \sigma_b^2) &= \int p_{\boldsymbol{x}},\, \boldsymbol{y} \,|\, \sigma_x^2,\, \sigma_b^2)\; d\boldsymbol{x} \\ &= \left(2\pi\sigma_b^2\right)^{-N/2} \alpha^{M/2} \left|\mathbf{D}^T\mathbf{D}\right|^{1/2} \left|\mathbf{A}^T\mathbf{A} + \alpha\mathbf{D}^T\mathbf{D}\right|^{-1/2} e^{-\mathcal{S}(\alpha)/2\sigma_b^2} \,. \end{aligned}$$

2. This strategy has no serious basis. Good asymptotic properties (when $N \to \infty$) are often mentioned for these estimators *without bias and with minimum variance*, but an estimator such as (3.10) also converges towards $\boldsymbol{x}_0$ in the same conditions, and faster, since for any finite N, its MSE is smaller.

By switching to logarithms, we obtain the log-marginal likelihood:

$$L(\alpha, \sigma_b^2) = \frac{M}{2}\log\alpha - \frac{N}{2}\log(2\pi\sigma_b^2) + \frac{1}{2}\log\left|\mathbf{D}^T\mathbf{D}\right| - \frac{1}{2}\log\left|\mathbf{A}^T\mathbf{A} + \alpha\mathbf{D}^T\mathbf{D}\right| - \frac{\mathcal{S}(\alpha)}{2\sigma_b^2}\,.$$

If this likelihood is sufficiently peaked, we can then satisfy ourselves with finding the $(\widehat{\alpha},\, \widehat{\sigma}_b^2)$ pair that maximizes $L(\alpha,\, \sigma_b^2)$. We have:

$$\frac{\partial L}{\partial \sigma_b^2} = -\frac{N}{2\,\sigma_b^2} + \frac{\mathcal{S}(\alpha)}{2\,\sigma_b^4} = 0 \quad \Longrightarrow \quad \widehat{\sigma}_b^2 = \frac{\mathcal{S}(\alpha)}{N}\,,$$

the"usual" estimator for variance. It is, however, difficult to maximize L as a function of α. We will thus content ourselves with finding $\widehat{\alpha}$ by exploring a discrete grid, since the result $\widehat{\boldsymbol{x}}(\alpha)$ is, in general, sensitive only to variations of the order of magnitude of α [FOR 93, THO 91].

3.8.3. *Wiener filtering*

The "linear-quadratic" framework is the only one that allows a statistical interpretation to be given in the *infinite dimension* problem [FRA 70]:

$$y = A\,x + b\,, \qquad x \in \mathcal{X}\,,\ \ y \in \mathcal{Y}\,. \tag{3.12}$$

For this, we assume that the functions x, y and b appearing in equation (3.12) are particular *trajectories* or *realizations*, of *stochastic processes* X, Y and B, linked by an analog relation[3]:

$$Y = A\,X + B\,. \tag{3.13}$$

If the zero-mean process X depends on a variable r, its *covariance function* is defined as $\Gamma_X(r,\, r') = \mathrm{E}(X(r)\,X(r'))$, and we assume that the functions x, trajectories of the process X, belong to a Hilbert space $\mathcal{X}$ and the functions y and b, the respective trajectories of Y and B, belong to the same Hilbert space $\mathcal{Y}$ (which may be distinct from $\mathcal{X}$). The covariance (function) of X can thus be considered as the kernel of an operator R_X defined on the space $\mathcal{X}$:

$$(R_X\,\phi)(r) = \int \Gamma_X(r,\, r')\,\phi(r')\;dr', \qquad \phi \in \mathcal{X}\,.$$

The inverse problem is to estimate a realization x of X, given the observation data of the realization y of Y and probabilistic prior knowledge on the processes X and B.

3. Here, for the sake of simplicity, we also assume that processes X, Y and B have zero mean. This hypothesis is not restrictive as, if they do not, the processes can always be centered and, thanks to the linearity of A, relation (3.13) remains true for the centered processes.

In the special case where X is a Gaussian process (or any linear transformation – such as the derivative – of a Gaussian process), the *a priori* probability law for X can be written symbolically[4]:

$$p_X(x) \propto \exp\left\{-\frac{1}{2}\,\langle x, R_X^{-1}x\rangle_{\mathcal{X}}\right\}\,.$$

If we take the hypothesis that the noise process B is additive, white and Gaussian of variance σ^2, the *a posteriori* law can be written:

$$p_X(x\,|\,Y=y) \propto \exp\left\{-\frac{1}{2\sigma^2}\,\left(\|y-A\,x\|_{\mathcal{Y}}^2+\sigma^2\,\langle x, R_X^{-1}x\rangle_{\mathcal{X}}\right)\right\}\,.$$

The best estimator of x, given the observation data y, depends on the choice of the optimality criterion but, in this case, if we choose the maximum of the *a posteriori* law or the MSE and if we factorize the covariance operator according to:

$$R_X = (C^*C)^{-1}, \tag{3.14}$$

the solution minimizes the criterion $\|y-Ax\|_{\mathcal{Y}}^2+\sigma^2\,\|Cx\|_{\mathcal{X}}^2$. It follows that $\widehat{x}=Gy$, where G is given by (2.7) with $\alpha=\sigma^2$. Moreover, if we define the operator $R_B=\sigma^2\,Id$, where Id is the identity operator in $\mathcal{Y}$ (R_B is the covariance operator of white noise), then G can also be written in the form:

$$G = R_X\,A^*(A\,R_X\,A^*+R_B)^{-1}, \tag{3.15}$$

which is the form of the *Wiener filter*. Put differently, the Tikhonov regularizer (2.7) is analogous to a Wiener filter in the case of white noise, provided that the constraint operator C that appears in it is linked to the covariance operator R_X by relation (3.14). Note, however, that second order ergodic processes have trajectories of finite power but infinite energy: $\mathcal{X}$ is not a summable square function space.

Equation (3.15) differs from the usual expression for a Wiener filter expressed in the Fourier domain. In fact, the expression above is more general. We will find the usual formulation again in Chapter 4, by taking advantage of additional hypotheses such as the convolutional structure of operator A and the weak stationarity (second order) of random processes X and B.

In contrast, in the case of non-quadratic functionals $\mathcal{G}$ or $\mathcal{F}$, the minimization of criterion (2.5) does not have a systematic statistical interpretation. In substance, the difficulty comes from the fact that the mathematical quantity characterizing the probability of a random process indexed on a space of finite dimension is, in this case, a set of functions having no direct relation with (2.5) and not allowing the likelihood function to be defined naturally.

4. In fact, the law of a process is given by the joint law of the n random variables $X(r_1)$, $X(r_2)$, …, $X(r_n)$, $\forall n \in \mathbb{N}$, $\forall (r_1, r_2, \ldots, r_n) \in \mathbb{R}^n$.

3.9. Bibliography

[BER 94] BERNARDO J. M., SMITH A. F. M., *Bayesian Theory*, Wiley, Chichester, UK, 1994.

[BES 86] BESAG J. E., "On the statistical analysis of dirty pictures (with discussion)", *J. R. Statist. Soc. B*, vol. 48, num. 3, p. 259-302, 1986.

[BRO 86] BROWN L. D., *Foundations of Exponential Families*, vol. 9, Hayward, CA, IMS Lecture Notes, Monograph Series edition, 1986.

[COX 61] COX R., *The Algebra of Probable Inference*, Johns Hopkins University Press, Baltimore, MD, 1961.

[DEM 77] DEMPSTER A. P., LAIRD N. M., RUBIN D. B., "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Statist. Soc. B*, vol. 39, p. 1-38, 1977.

[DEM 89] DEMOMENT G., "Image reconstruction and restoration: overview of common estimation structure and problems", *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-37, num. 12, p. 2024-2036, Dec. 1989.

[FOR 93] FORTIER N., DEMOMENT G., GOUSSARD Y., "GCV and ML methods of determining parameters in image restoration by regularization: fast computation in the spatial domain and experimental comparison", *J. Visual Comm. Image Repres.*, vol. 4, num. 2, p. 157-170, June 1993.

[FRA 70] FRANKLIN J. N., "Well-posed stochastic extensions of ill-posed linear problems", *J. Math. Anal. Appl.*, vol. 31, p. 682-716, 1970.

[GAS 92] GASSIAT E., MONFRONT F., GOUSSARD Y., "On simultaneous signal estimation and parameter identification using a generalized likelihood approach", *IEEE Trans. Inf. Theory*, vol. 38, p. 157-162, Jan. 1992.

[GEM 84] GEMAN S., GEMAN D., "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, num. 6, p. 721-741, Nov. 1984.

[JAY 82] JAYNES E. T., "On the rationale of maximum-entropy methods", *Proc. IEEE*, vol. 70, num. 9, p. 939-952, Sep. 1982.

[JEF 39] JEFFREYS, *Theory of Probability*, Oxford Clarendon Press, Oxford, UK, 1939.

[KAS 94] KASS R. E., WASSERMAN L., Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography, Technical report no. 583, Department of Statistics, Carnegie Mellon University, 1994.

[LIT 83] LITTLE R. J. A., RUBIN D. B., "On jointly estimating parameters and missing data by maximizing the complete-data likelihood", *Amer. Statist.*, vol. 37, p. 218-220, Aug. 1983.

[MAR 87] MARROQUIN J. L., MITTER S. K., POGGIO T. A., "Probabilistic solution of ill-posed problems in computational vision", *J. Amer. Stat. Assoc.*, vol. 82, p. 76-89, 1987.

[POL 92] POLSON N. G., "On the expected amount of information from a non-linear model", *J. R. Statist. Soc.*, vol. 54, num. B, p. 889-895, 1992.

[ROB 97] ROBERT C. P., *The Bayesian Choice. A Decision-Theoretic Motivation*, Springer Texts in Statistics, Springer Verlag, New York, NY, 1997.

[TAR 87] TARANTOLA A., *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier Science Publishers, Amsterdam, The Netherlands, 1987.

[THO 91] THOMPSON A., BROWN J. C., KAY J. W., TITTERINGTON D. M., "A study of methods of choosing the smoothing parameter in image restoration by regularization", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-13, num. 4, p. 326-339, Apr. 1991.

[TIT 85] TITTERINGTON D. M., "Common structure of smoothing techniques in statistics", *Int. Statist. Rev.*, vol. 53, num. 2, p. 141-170, 1985.