Regularization Methods and Inverse Problems: An Information Theory Standpoint

Jérôme IDIER, Ali MOHAMMAD-DJAFARI, and Guy DEMOMENT
Laboratoire des signaux et systèmes (CNRS/ESE/UPS)
Supélec, Plateau de Moulon, 91192 GIF-SUR-YVETTE Cedex, France

ABSTRACT

In a number of engineering topics we are faced with the inverse problem of recovering the spatial distribution of some scalar or vector quantity from measurements of the interaction of an investigated medium with an incident wave. The common feature of such image reconstruction problems is that they are often ill-posed or ill-conditioned. We review first the basic aspects of standard regularization theory. Then, using an information-based approach, we show that existing regularization criteria, which were introduced in the literature using very different approaches, can be interpreted as special cases of an entropy, in spite of their apparent variety. Finally, we discuss its limitations and present the Bayesian statistical approach which allows local properties to be introduced in the estimated image through Markov random fields and associated local energy functions.

1 ILL-POSED PROBLEMS

This paper deals with methods for solving inverse problems, and more specifically with their links to information theory. In these problems, the object of interest cannot be observed directly and must be determined from the observed data in order to get rid of the defects of the observation device.

The object $\boldsymbol{x}$ and the measurements $\boldsymbol{y}$ are related through an equation $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{y}) = 0$ which can often be solved for $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x})$. Computing $\boldsymbol{y}$ from $\boldsymbol{A}$ and $\boldsymbol{x}$ is a *direct problem*. Conversely, evaluation of $\boldsymbol{x}$ knowing $\boldsymbol{A}$ and $\boldsymbol{y}$ is the *inverse problem*. Of course no experimental device is completely free from uncertainties whose simplest cause is the finite precision of measurements. It is therefore more realistic to consider that the unknown object and the measurements are related through a relation of the form: $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}) \diamond \boldsymbol{n}$ where $\diamond \boldsymbol{n}$ accounts for the degradation induced on the ideal representation $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x})$ by a process $\boldsymbol{n}$ referred to as *noise*. When the observation mechanism can be approximated by a linear transformation corrupted by additive noise, the previous relation reduces to

$$\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{n}. \tag{1}$$

The scope of the paper is limited to inverse problems that can be modeled in this form. In spite of the restrictive character of this model, the corresponding inverse problem is generic in the sense that its resolution can give rise to several other methods (see, for instance, (?, 18)).

As the sizes of $\boldsymbol{x}$ and $\boldsymbol{y}$ are not necessarily the same, a natural idea to solve the problem consists of minimizing a least squares criterion of the form:

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{G}(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x}) = \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x}\|^2 . \tag{2}$$

The least squares or minimum norm solution can be written as: $\widehat{\boldsymbol{x}}_0 = (\boldsymbol{A}^t\,\boldsymbol{A})^{-1}\,\boldsymbol{A}^t\,\boldsymbol{y}$ or $\widehat{\boldsymbol{x}}_0 = \boldsymbol{A}^\dagger\,\boldsymbol{y}$, according to whether the normal matrix is regular or whether it only admits a generalized inverse $\boldsymbol{A}^\dagger$. This seems to be a reasonable choice from a statistical standpoint at least, as $\widehat{\boldsymbol{x}}_0$ is, under our assumptions, an unbiased and minimum variance solution. However this solution is generally unacceptable because $\boldsymbol{A}$ is ill-conditioned: the resulting amplification of noise is beyond any acceptable level (13, 2).

2 REGULARIZATION OF AN ILL-POSED PROBLEM

Several methods have been proposed to stabilize and solve ill-posed problems. A now classical way of reaching this goal, is provided by *regularization* (23, 19, 22). The basic idea consists in giving up the hope of obtaining an exact solution from imperfect data, in defining a class of *admissible solutions* $\{\boldsymbol{x} : \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x} \| \leq \| \boldsymbol{n} \|\}$, and in selecting, within this class, a solution that will be considered acceptable in the sense that it is consistent with some *prior information*. For this purpose, the solution $\widehat{\boldsymbol{x}}(\alpha, \boldsymbol{y})$ is often defined as the minimizer of a criterion such as

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{G}(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x}) + \alpha\,\mathcal{F}(\boldsymbol{x}) \qquad 0 < \alpha < +\infty . \tag{3}$$

This criterion is specifically designed for: (i) insuring, to some extent, the fidelity of the solution to the data (first component of the criterion), and (ii) favoring some desirable properties which summarize the prior knowledge of the solution (second component of the criterion).

Selection of functionals $\mathcal{F}$ and $\mathcal{G}$ is a qualitative choice which determines how regularization is performed. Conversely the choice of α, which is referred to as the regularization coefficient, is quantitative and controls the tradeoff between the two sources of information. A perfect fidelity to the data is obtained with $\alpha = 0$, whereas a total fidelity to the priors is obtained with $\alpha = \infty$. In the standard regularization approach of Phillips, Twomey and Tikhonov, $\mathcal{F}$ and $\mathcal{G}$ are both quadratic (25):

$$\mathcal{G}(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x}) = \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x}\|^2, \qquad \mathcal{F}(\boldsymbol{x}) = \| \boldsymbol{D}_k\,\boldsymbol{x} \|^2 . \tag{4}$$

This approach gave rise to important theoretical and applied studies. It can be extended to some cases in which the direct problem is nonlinear (22). However, the questions of the choice of *regularizing functional* $\mathcal{F}(x)$ and of *regularizing coefficient* α (also referred to as *hyper-parameter*) remain open.

3 REGULARIZATION, INFORMATION, AND ENTROPY

It can be shown that several classical regularization criteria are specific cases of *maximum entropy on the mean* (7, 1). In this problem, constraints on the object $\boldsymbol{x}$ are specified through a convex set $\mathcal{C}$ that it is supposed to belong to, and the solution is chosen as the mean value of the distribution which is the closest to a reference measure μ on $\mathcal{C}$, with respect to the Kullback distance. This approach provides a general framework for interpretation of these criteria which thereby appear as entropies whose form is directly connected to the prior information. As an example, the object could be constrained to belong to the following convex set $\mathcal{C} = \{x \in I\!R^N \,/\, x_k \in]a_k, b_k[,\ k = 1, \ldots, N\}$, where a_k, b_k are known constants.

3.1 Information principle

Let us start with the ideal noiseless case. The a priori information plays an important part as the inversion process starts with specification of convex set $\mathcal{C}$ and of a reference measure $d\mu(\boldsymbol{x})$ defined over $\mathcal{C}$. Assume that the data $\boldsymbol{y}$ are obtained through application of matrix $\boldsymbol{A}$ onto the mean value of a process $\boldsymbol{X}$ under a probability distribution P defined over $\mathcal{C}$. Since $\mathcal{C}$ is convex, the mean value $\mathrm{E}_P\{\boldsymbol{X}\}$ with respect to P belongs to $\mathcal{C}$ and the convex constraint is necessarily satisfied. But since the data constraint $\boldsymbol{y} = \boldsymbol{A}\,\mathrm{E}_P\{\boldsymbol{X}\}$ is not sufficient to define a unique probability distribution P, an additional information theoretic principle must be utilized. For this purpose, we introduce the Kullback information measure, which, for a reference measure μ and a probability distribution P, is defined by

$$\mathcal{K}(P, \mu) = \int \log \frac{\mathrm{d}P}{\mathrm{d}\mu}\, \mathrm{d}P \tag{5}$$

when P is absolutely continuous with respect to μ, and infinite otherwise.

Therefore P_{ME} is chosen as the distribution which minimizes $\mathcal{K}(P, \mu)$ under the constraints "*on the mean*" $\boldsymbol{A}\,\mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{y}$. In other words, P_{ME} is the probability distribution which is the closest to μ, in the sense of the Kullback divergence, among those which match the data on the mean. It is well known that the solution to such an optimization problem, when it exists, belongs to the exponential family $dP_{\boldsymbol{s}}(\boldsymbol{x}) = \exp\{\boldsymbol{s}^t\boldsymbol{x} - \log Z(\boldsymbol{s})\}\ d\mu(\boldsymbol{x})$, in which the natural parameter, $\boldsymbol{s} = \boldsymbol{A}^t\boldsymbol{\lambda}$, is a function of the Lagrange multiplier $\boldsymbol{\lambda}$ of the constrained optimization problem. Z is the partition function and $\mathcal{F}^* = \log Z$ refers to the log-Laplace transform of measure $d\mu(\boldsymbol{x})$.

3.2 Dual problems

The theory of duality (16) indicates that the optimal solution to the previous problem is equal to the optimal value of its dual problem:

$$\underset{P \in \mathcal{P}_y}{Inf}\ \mathcal{K}(P, \mu) = \underset{\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\lambda}}}{Sup}\ \left\{\boldsymbol{\lambda}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{A}^t\boldsymbol{\lambda})\right\}, \tag{6}$$

where $\mathcal{P}_y = \{P : \boldsymbol{A}\,\mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{y}\}$ denotes the set of normalisable distributions which fulfill the constraint on the mean, and where $\mathcal{D}_{\boldsymbol{\lambda}}$ refers to the set $\{\boldsymbol{\lambda} \;:\; Z(\boldsymbol{A}^t\boldsymbol{\lambda}) < \infty\}$ which is very often equal to $I\!R^M$; in which case the dual problem is unconstrained. It is important to note that the dual criterion $\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{A}^t\boldsymbol{\lambda})$ is strictly convex by construction.

If we note by $\boldsymbol{m} = \int \boldsymbol{x} d\mu(\boldsymbol{x})$, it is possible to show that

$$\widehat{\boldsymbol{x}} = E_{P_{ME}}\{\boldsymbol{X}\} = \underset{\boldsymbol{x}\in\mathcal{C}_{\boldsymbol{y}}}{arg\,min}\, \mathcal{F}(\boldsymbol{x}), \quad \text{with} \quad \mathcal{C}_{\boldsymbol{y}} = \{\boldsymbol{x} : \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{y}\}, \tag{7}$$

and

$$\mathcal{F}(\boldsymbol{x}) = \underset{P\in\mathcal{P}_{\boldsymbol{x}}}{Inf}\; \mathcal{K}(P,\mu) = \underset{\boldsymbol{\lambda}\in\mathcal{D}_{\boldsymbol{\lambda}}}{Sup}\; \left\{\boldsymbol{\lambda}^t\boldsymbol{x} - \mathcal{F}^*(\boldsymbol{\lambda})\right\}. \tag{8}$$

This equation shows that $\mathcal{F}$ is the convex conjugate of $\mathcal{F}^*$ (and also the *Cramèr transform* of μ as $\mathcal{F}^*$ is the log-Laplace transform of μ). The major interesting properties of this transform in our problem are the following:
– $\mathcal{F}$ is continuously differentiable and strictly convex on $\mathcal{C}$,
– $\mathcal{F}(\boldsymbol{x}) = \infty$ for $\boldsymbol{x} \notin \mathcal{C}$ and its derivative takes infinite values on the boundary of $\mathcal{C}$,
– $\mathcal{F}(\boldsymbol{x}) \geq 0$ and the equality is reached when $\boldsymbol{x} = \boldsymbol{m}$.

This possibility to switch between primal problems is known as the "contraction principle" in statistical physics. From this standpoint, functional $\mathcal{F}$ can be considered as an entropy measure. Strict convexity results in simple implementation and guarantees the uniqueness of the solution. The second properties shows that descent methods yield a solution that belongs to $\mathcal{C}$, even if constraint $\boldsymbol{x} \in \mathcal{C}$ is not explicitly specified within the algorithm. The last property shows that $\mathcal{F}$ can be considered as a distance measure between $\boldsymbol{x}$ and $\boldsymbol{m}$. A few examples are given hereafter.

3.3 Examples

When there is no specific constraint on the object, then $\mathcal{C} = I\!R^n$. If a Gaussian distribution is chosen as the reference measure μ on $\mathcal{C}$, it can be easily shown that the Cramèr transform $\mathcal{F}$ is a simple quadratic regularization functional whose connection with Gaussian priors has already been pointed at. Assume now that the object is positive and that the reference distribution is a Poisson distribution assumed to be separable. The entropy $\mathcal{F}$ becomes a generalized form of Shannon entropy, but, if the reference measure is an exponential distribution, the definition of the Cramèr transform yields the Burg entropy.

In this manner, all regularizing functional of classical regularization theory can be derived; new forms, either explicit or implicit, can also be obtained, thanks to dual analysis (1).

3.4 Noisy data

The same approach can be used in order to account for the observation noise, regardless of whether it is characterized by a Gaussian distribution or not (1). One way to do this is to introduce an *extended object* $\tilde{\boldsymbol{x}}^t = [\boldsymbol{x}^t,\ \boldsymbol{n}^t]$, and rewrite a direct problem under the form $\boldsymbol{y} = \tilde{\boldsymbol{A}}\,\tilde{\boldsymbol{x}}$, with $\tilde{\boldsymbol{A}} = [\boldsymbol{A},\ \boldsymbol{1}]$. Vector $\tilde{\boldsymbol{x}}$ belongs to convex set $\tilde{\mathcal{C}}$ in IR^{N+M} (where M is the dimension of $\boldsymbol{b}$); $\tilde{\mathcal{C}}$ can be factorised as the convex product of $\mathcal{C}$ and of $\mathcal{B}$, $\tilde{\mathcal{C}} = \mathcal{C} \times \mathcal{B}$, where $\mathcal{B}$ denotes the convex hull of noise $\boldsymbol{n}$. We then use a reference measure ν over the noise set.

Now if we assume that the object and the noise process are independent, we obtain $\tilde{\mu} = \mu \otimes \nu$ and the entropy, which is the Cramèr transform of $\tilde{\mu}$, can simply be written as $\mathcal{F}_{\tilde{\mu}}(\tilde{\boldsymbol{x}}) = \mathcal{F}_{\mu}(\boldsymbol{x}) + \mathcal{F}_{\nu}(\boldsymbol{n})$. Estimation of the extended object can be carried out through constrained minimization of $\mathcal{F}_{\tilde{\mu}}(\tilde{\boldsymbol{x}})$, where the constraint is defined by $\boldsymbol{y} = \tilde{\boldsymbol{A}}\,\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$. It therefore melts down to unconstrained minimization of the compound criterion

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_{\tilde{\mu}}([\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}]^t) = \mathcal{F}_{\mu}(\boldsymbol{x}) + \mathcal{F}_{\nu}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}). \tag{9}$$

Therefore, specific noise distribution can be accounted for without losing the attractive properties of this criteria: global criterion (9) remains convex, and the convex constraint is automatically satisfied. When the noise is Gaussian, it can be shown that this yields the minimization of compound criterion $\mathcal{J}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{x}) + \alpha \parallel \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\parallel^2$, which has been introduced in the previous chapter through heuristic considerations.

Maximum entropy on the mean therefore provides an explanation of the generic form of classical regularized criteria, through duality techniques it also yields efficient solutions. However, with the important exception of the Gaussian measure, this approach can be applied to separable measures only, and it is not possible to introduce local a priori information on the object, such as correlations or contours. In addition, the question of the choice of the regularization coefficient, or more generally of the parameters which control the reference measure, is left unanswered in this approach. The solutions provided by Bayesian approach are examined in the next section.

4 BAYESIAN APPROACH TO REGULARIZATION

There are at least two reasons for casting inverse problem solving into a Bayesian framework. First, it allowed the development of local energy functions and Markov models that have had a long lasting influence on low level image processing. Second, it offers the most consistent and complete answers to the problem of the choice of hyper-parameter values.

4.1 Local energy functions

The Gaussian distributions associated with linear direct models yields linear estimators, and therefore very convenient algorithmic structures. However they can only capture very simplistic information that are basically limited to second order characteristics. A maximum entropy distribution retains part of the properties of Gaussian priors such as the convexity of the regularized criterion, while accounting for other statistical distribution of the noise process.

Expressing the local image properties in a quantitative manner (homogeneous regions separated by contours for example) can be carried out in the general framework of energy functions which were introduced in image processing by considering that the object $\boldsymbol{x}$ is, like the noise, a realization of a random variable (11, 3). In a preliminary step, one must define a neighborhood system $\{d_i\}$ in which d_i denotes the set of pixels which are assumed to interact directly with pixel i. Assume, for example, that values x_i be quantified. A very common characteristic of images is that the intensity values of neighboring sites are likely to be similar. Then one defines local energies

$$V(x_i, x_j) = \begin{cases} -1 & x_i = x_j \quad j \in d_i \\ +1 & x_i \neq x_j \quad j \in d_i \\ 0 & else \end{cases} \tag{10}$$

These interaction energies are evaluated over all neighboring pixels and then added up to define the energy of the image

$$\mathcal{V}(\boldsymbol{x}) = \sum_i \sum_j V(x_i, x_j) = \sum_k V_k(\boldsymbol{x})\,.$$

V_k denotes the energy associated with the k^{th} set of neighboring pixels, and k varies over the set of all pairs of neighboring sites. The general idea of these derivations is that $\mathcal{V}(\boldsymbol{x})$ must be small for images that fulfill the properties used to define the priors, in this example images whose neighboring pixels tend to have similar intensity values. However energy functions can also express other properties such as the existence of almost uniform regions separated by clear discontinuities. In order to model these discontinuities, local energy functions have been recently introduced. The original image $\boldsymbol{x}$ is considered as a pair $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{t})$ in which a vector of pixel intensities $\boldsymbol{z}$ which could be observed directly with a perfect device, is associated with a vector $\boldsymbol{t}$ of additional hidden variables which express non-observable characteristics such as the presence of a contour, of a texture etc... Then the global energy is defined as

$$\mathcal{F}(\boldsymbol{x}) = \mathcal{V}_{\boldsymbol{z}}(\boldsymbol{z}) + \mathcal{V}_{\boldsymbol{t}}(\boldsymbol{t}) + \mathcal{V}_{\boldsymbol{z}\boldsymbol{t}}(\boldsymbol{z}, \boldsymbol{t})\,,$$

which is made up of three terms: an intensity term $\mathcal{V}_{\boldsymbol{z}}(\boldsymbol{z})$, a contour term $\mathcal{V}_{\boldsymbol{t}}(\boldsymbol{t})$ and a third term $\mathcal{V}_{\boldsymbol{z}t}(\boldsymbol{z}, \boldsymbol{t})$ which describes the interactions between contours and pixel values (11).

4.2 Bayesian approach to regularization

An important proportion of statistical inference methods is based upon the use of a priori information on quantities to be estimated. It is not surprising that such techniques presents such tight connections with the regularization principles presented in chapter 2. In a Bayesian context, the a priori information on object $\boldsymbol{x}$ is expressed in the form of an a priori probability distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$. Bayes rule allows us to combine these priors with the information contained in the observed data so as to obtain the a posteriori probability distribution

$$p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{A},\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{A},\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{A},\boldsymbol{\theta})}\,. \tag{11}$$

In the above equation, $\boldsymbol{\theta}$ is a vector of *hyper-parameters* which is made of the parameters of the prior distributions of the errors and of the object, and $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{A},\boldsymbol{\theta})$ denotes the probability distribution of the data conditioned on $\boldsymbol{x}$. It is completely determined by the knowledge of the direct model (1) and of the probability distribution of the noise. The last term is a normalization factor of the posterior distribution:

$$p(\boldsymbol{y}|\boldsymbol{A},\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{A},\boldsymbol{\theta})\, p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}\,. \tag{12}$$

In the Bayesian standpoint, (11) is the solution to the inversion problem, as it sums up all available information on $\boldsymbol{x}$. The manipulation of probability distribution is generally tedious and often even impossible; this is why a decision must be made of each object component. A very common choice consists of selecting for $\boldsymbol{x}$ a value which maximizes the posterior distribution

$$\widehat{\boldsymbol{x}}_{MAP} = \underset{\boldsymbol{x}}{arg\,max}\, p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{A},\boldsymbol{\theta})\,. \tag{13}$$

This is only one of many possible solutions. The MAP estimations corresponds to minimization of an average cost in which the cost function presents a zero/one shape. In the context of Markov based image modeling, other cost functions have recently attracted the interest of investigators. They yield maximizations of marginal probability distributions (3, 17).

In the framework of this study, *i.e.* finite dimension inverse problems, it is clear that regularization according to the general principle presented in chapter 3, which yields minimization of a criterion such as (3), is equivalent to choosing

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{A},\boldsymbol{\theta}) \propto \exp\{-(1/2\,\sigma^2)\,\mathcal{G}(\boldsymbol{y}-\boldsymbol{A}\,\boldsymbol{x})\}\,.\ \text{and}\ p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \exp\{-(\alpha/2\,\sigma^2)\,\mathcal{F}(\boldsymbol{x})\}\,. \tag{14}$$

We only have a Bayesian interpretation of regularization methods; here we will not discuss whether this interpretation represents a justification of such methods or not. We only recall that most local energy functions used in the image restoration fields have been introduced in the Bayesian framework: through (14)

these energy functions define $\boldsymbol{x}$ as a random Markov field (11, 3). In addition, the Bayesian interpretation provides new ways of deriving hyper-parameter determination methods. In order to implement all methods described in the previous chapters, it is necessary to choose the value of regularization parameter α and, more generally, of all parameter $\boldsymbol{\theta}$ which defines distance measures $\mathcal{F}$ et $\mathcal{G}$: the noise variance, the correlation parameters of the object, and the parameters of the local energy functions. Determination of $\boldsymbol{\theta}$ is the most critical step of image reconstruction and restoration methods. Though this problem is not yet solved in a completely satisfactory manner, Bayesian approach provides adequate tools to tackle it.

4.3 Criteria selection

In image modeling, maximum entropy distributions can be naturally generalized under the form of Markov fields, which capture structural information connected to local properties of the objects. Indeed, the possibility of describing homogeneous area separated by sharp discontinuities is crucial in image processing (14). However when Markov representations associated with Gibbs energy functions are used, the major problem lies in the difficulty of evaluating the posterior distribution of the object, because of the non-linearities of the energy functions and of the very large number of possible object configurations. For example, marginal prior distribution of the object $p(x_i)$ cannot be evaluated; only conditional distribution $p(x_i|x_j, j \in d_i)$ can be computed. This difficulty has been partly alleviated by *Monte Carlo* type stochastic techniques (Gibbs sampler and Simulated annealing). But, in general, the amount of computation grows rapidly with the dimension of the neighborhood system, which makes the procedure very impractical in inverse problems where matrix $\boldsymbol{A}$ is not a local operator. The ICM algorithm (3) is a determistic version of simulated annealing, it presents faster convergence, but may get stuck in a local minimum of the regularized criterion.

In order to overcome this difficulty, a graduated non-convexity (GNC) optimization procedure without locality constraint may also be employed. This is a deterministic sub-optimal relaxation method initially introduced in computer vision (4). The idea consists of designing a series $\mathcal{V}^{(n)}(\boldsymbol{x})$ of energy functions that converges to $\mathcal{V}(\boldsymbol{x})$ such that: (i) the initial function is convex and (ii) the series of local minima associated with $\mathcal{V}^{(n)}(\boldsymbol{x})$ converges toward the global minimum of $\mathcal{V}$. Solutions to some inverse problems obtained with this technique are very encouraging (20, 6).

Finally, since Markov fields with non-convex energy functions yield difficult global optimization problems, one may prefer to use generalized Gaussian Markov fields, which are similar to models used in robust estimation and which yield convex criteria (5, 6).

4.4 Choice of hyper-parameter values

There are relatively few methods for the determination of hyper-parameters (24). When they are limited to a unique regularisation coefficient α, and when the regularizing functional $\mathcal{F}(\boldsymbol{x})$ is quadratic, *cross-validation methods* provides acceptable solutions (12). This method presents interesting asymptotic properties, but has a clear justification only in the case of standard regularization methods with quadratic criteria.

Hyper-parameters $\boldsymbol{\theta}$ make up a second description level for the problem that is essential for "stiffening" the first level, *i.e.* the parameters or the object $\boldsymbol{x}$. In inverse problem solving, adequate specification of hyper-parameter values is important in order to obtain an acceptable solution. However these values do not present any particular significance. In a Bayesian approach, one may then separate two inference levels. The first one infers on $\boldsymbol{x}$ for a given value of $\boldsymbol{\theta}$ through the posterior distribution of equation (11). The second one infers on $\boldsymbol{\theta}$ thanks to a similar relationship:

$$p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{A}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{A})\, p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{A})}{p(\boldsymbol{y}|\boldsymbol{A})}\,. \tag{15}$$

It should be underlined that the likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{A})$ in the second level is equal to the normalizing factor of the posterior distribution in the first level. This is a specific characteristic of the Bayesian approach.

In the very common case when this term is "spiky" enough, which means that the data $\boldsymbol{y}$ contains enough information, the influence of the prior distribution $p(\boldsymbol{\theta}|\boldsymbol{A})$ can be neglected and the second inference level can be solved though maximization of this likelihood function. However one must now solve a marginalization problem:

$$p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{A}) = \int p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{A})\, d\boldsymbol{x} = \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{A})\, p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}\,. \tag{16}$$

Such an expression rarely yields an explicit solution.

To avoid the difficulty, one may introduce *hidden variables* $\boldsymbol{z}$ that complete observations $\boldsymbol{y}$ in order to simplify the evaluation of the new likelihood function $p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{A})$. It is then necessary to maximize conditional expectations using iterative techniques that may be either deterministic or stochastic (EM and SEM algorithms) (9). Stochastic techniques were introduced in order to overcome the difficulties in the maximization in the likelihood functions using classical optimization techniques.

It should also be noticed that the *generalized likelihood function*

$$p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{A}) = p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{A})\, p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{A}) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{A})\, p(\boldsymbol{x}|\theta) \tag{17}$$

sums up all the information related to the first inference level. One may think of performing the joint maximization of this generalized likelihood function with respect to $\boldsymbol{x}$ and $\boldsymbol{\theta}$, thereby avoiding the integration problem raised by (16). For a given value of $\boldsymbol{\theta}$, the maximum generalized likelihood (MGL) is equal to the MAP. However, for a given value of $\boldsymbol{x}$, the situation is much less favorable: in general, the generalized likelihood function is not upper bounded over its domain definition, and local maxima may not even exist (10).

5 CONCLUSION

The Bayesian approach offers remarkably coherent framework for solving inference problems in the difficult situation when several information sources are available: experimental data, direct model, prior information. The limitations are well known and are essentially of a computational nature. Clearly, it is very unlikely for this approach to yield satisfactory results for a complex problem if the models are simplistic and if the optimization methods are too coarse. This is probably why the Gaussian based initial attempt performed about twenty years ago in the areas of image restoration and noise filtering were not that successful. However, in the area of computer vision and since the milestone paper of Geman and Geman in 1984, these techniques have been revived in a spectacular manner. They associate Markov fields, Gibbs sampling, simulated annealing, parallel computing in problems like image segmentation, edge detection, texture extraction etc... These works have provided a very strong theoretical basis to these domains and have a growing influence on the larger area of inverse problems.

It is our belief that this avenue is sill widely open for investigation, particularly on subjects like the design of probabilistic models well-suited to the problem to be solved, determination of hyper-parameter values or derivation of optimization techniques. This research is presently conducted along the following lines.
– *Forward problem modeling.* Imaging problems are very often non-linear by nature. This is particularly the case for diffraction tomography and conductivity imaging. Therefore non-linear direct problems will have to be introduced in these areas.
– *Modeling of the studied objects.* Prior Markov models are well suited to the representation of images that present discontinuities. However this raises difficulties with respect to the situation of computer vision because of the non-local character of linear operators of the direct problem used in image reconstruction or restoration.
– *Choice of criteria.* In the Bayesian approach, inversion corresponds to the determination of an *a posteriori* distribution. However, since complete evaluation of such a distribution cannot be practically carried out, one generally reverts to the use of a point estimator, which is generally a MAP estimator. It is therefore important to evaluate the consequences of such a choice and to propose other solutions if necessary. Similarly, the question of the choice of a criterion for hyper-parameter

estimation is still widely open.
– *Algorithms for criterion optimization.* Once the criteria are chosen, their optimization often presents difficulties since they generally exhibit several local optima. It is therefore important to keep developing stochastic optimization algorithms, particularly within the "*expectation maximization*" (EM) family.

REFERENCES

(1) J.-F. Bercher, G. Le Besnerais, and G. Demoment, "Building convex criteria for solving linear inverse problems," *Proc. Intern. Workshop on Inverse Problems*, Ho-Chi-Minh City, 17-19 Jan. 1995, pp. 33-44.

(2) M. Bertero, T.A. Poggio, and V. Torre, "Ill-posed problems in early vision," *Proc. IEEE*, vol. 76, pp. 869-889, Aug. 1988.

(3) J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statist. Soc.* ser. B, vol. 48, pp. 259-302, 1986.

(4) A. Blake, and A. Zisserman, *Visual Reconstruction*, Cambridge, MA: MIT Press, 1987.

(5) C. Bouman, and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Processing*, vol. 2, pp. 296-310, Jul. 1993.

(6) H. Carfantan, and A. Mohammad-Djafari, "A Bayesian approach for nonlinear inverse scattering tomographic imaging," *Proc. IEEE ICASSP'95*, Detroit, May 1995, pp. 2311-2314.

(7) D. Dacunha-Castelle, and F. Gamboa, "Maximum d'entropie et problème des moments," *Ann. Instit. Henri Poincaré*, vol. 26, pp. 567-596, 1990.

(8) G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 2024-2073, Dec. 1989.

(9) A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, ser. B, vol. 39, pp. 1-38, 1977.

(10) É. Gassiat, F. Monfront, and Y. Goussard, "On simultaneous signal estimation and parameter estimation using a generalized likelihood approach," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 157-162, 1992.

(11) S. Geman, and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721-741, Nov. 1984.

(12) G. Golub, M. Heath, and G. Wahba, "Generalized Cross-Validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215-223, 1979.

(13) G.T. Herman, H.K. Tuy, H. Langenberg, and P.C. Sabatier, *Basic Methods of Tomography and Inverse Problems*, Adams Hilger, 1987.

(14) J. Idier, and Y. Goussard, "Markov modeling for Bayesian restoration of two-dimensional layered structures," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1356-1373, Jul. 1993.

(15) K.C. Li, "Asymptotic optimality of C_L and GCV in ridge regression with application to spline smoothing," *Ann. Statist.*, vol. 14, pp. 1101-1112, 1986.

(16) D. Luenberger, *Optimization by vector space methods*, New York: Wiley, 1969.

(17) J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Statist. Ass.*, vol. 82, pp. 76-89, 1987.

(18) T. Martin, and J. Idier, "A Bayesian non-linear inverse approach for electrical impedance tomography," *2nd Intern. Conf. Inverse Problems in Engng.*, Le Croisic, June 1996.

(19) M.Z. Nashed, "Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory," *IEEE Trans. Antennas Propag.*, vol. AP-29, pp. 220-231, 1981.

(20) M. Nikolova, A. Mohammad-Djafari, and J. Idier, "Inversion of large-support ill-conditioned linear operators using a Markov process with line process," *Proc. IEEE ICASSP'94*, Adelaide, Avril 1994, pp. 357-360.

(21) S.J. Reeves, and R.M. Mersereau, "Optimal estimation of the regularization parameter and stabilizing functional for regularized image restoration," *Opt. Engng.*, vol. 29, pp. 446-454, 1990.

(22) A. Tarantola, *Inverse problem theory: Methods for data fitting and model parameter estimation*, Amsterdam: Elsevier Science Publishers, 1987.

(23) A. Tikhonov, and V. Arsenin, *Solutions of Ill-Posed Problems*, Washington, DC: Winston and Sons, 1977.

(24) A.M. Thomson, J.C. Brown, J.W. Kay, and D.M. Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *Trans. IEEE Pattern Anal. Mach. Intell.*, vol. PAMI-13, pp. 326-339, 1991.

(25) D.M. Titterington, "Common structures of smoothing techniques in statistics," *Int. Statist. Rev.*, vol. 53, pp. 141-170, 1985.