BUILDING CONVEX CRITERIA FOR SOLVING LINEAR INVERSE PROBLEMS

Jean-François BERCHER, Guy LE BESNERAIS,

and

Guy DEMOMENT
Laboratoire des signaux et systèmes (CNRS-ESE-UPS)
Plateau de Moulon, 91190 Gif-sur-Yvette, France

ABSTRACT

In this paper we address the problem of building convenient criteria to solve linear and noisy inverse problems of the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$. Our approach is based on the specification of constraints on the solution $\boldsymbol{x}$ through its belonging to a given convex set $\mathcal{C}$. The solution is chosen as the mean of the distribution which is the closest to a reference measure μ on $\mathcal{C}$ with respect to the Kullback divergence, or cross-entropy. This is therefore called the Maximum Entropy on the Mean Method (MEMM). This problem is shown to be equivalent to the convex one $\boldsymbol{x} = \arg\min_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x})$ submitted to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ (in the noiseless case). Many classical criteria are found to be particular solutions with different reference measures μ. The MEMM also takes advantage of a dual formulation to exhibit dual problems, often unconstrained, whereas the direct problem is constrained and may not be explicit. The presence of additive noise is also integrated in the MEMM scheme, the object and noise being both searched for in an appropriate convex $\mathcal{C}'$ including the previous one $\mathcal{C}$. The MEMM then gives an unconstrained criterion of the form $\mathcal{F}(\boldsymbol{x}) + \mathcal{G}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$.

1. Problem statement and present answers

A general problem arising in experimental data processing is to estimate an object from a set of measurements. No experimental device, even the most elaborate, is entirely free from uncertainty. The simplest example is the finite working precision of the recording device and observations are also usually corrupted by noise. If we let $\boldsymbol{x}$ be the object, and $\boldsymbol{y}$ be the data vector, the object and the measurements are then related by a relation of the form $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}) \diamond \boldsymbol{n}$. In this general expression, $\boldsymbol{A}$ is a (non-)linear operator describing the essential part of the experiment, and the operation $\diamond \boldsymbol{n}$ accounts for the degradation of this ideal representation by a random process $\boldsymbol{n}$. In this communication, we will only consider the situations where the distortion mechanism can be correctly modeled as a *linear*

transformation of $\boldsymbol{x}$ and the addition of noise, so that the previous relation reduces to:

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

This is often the situation encountered in image reconstruction and restoration[4].

A natural idea for inverting Eq. (1) is to use the generalized inverse of $\boldsymbol{A}$. Unfortunately, it is generally ill-conditioned and the reconstruction suffers from an excessive amplification of any observation noise[4]. A (now) classical way of handling this kind of difficulty is *regularization* theory. The basic idea is to renounce the hope of obtaining an exact solution from imperfect data, and to define a class of *admissible solutions* by adding to Eq. (1) some extra *prior* information concerning what may be considered as a reasonable "physical" solution.

In this paper we are especially interested in the reconstruction of objects known to belong to some specified convex set. Examples of this situation are plentiful, let us only cite the problem of imaging *positive* intensity distributions, which arises in spectral analysis, astronomy, spectrometry, etc...In other specific problems, such as crystallography or tomography, lower and upper bounds on the image are known — and have to be taken into account in the reconstruction process. Such constraints may be specified by the belonging of the object to the convex set

$$\mathcal{C} = \{\ \boldsymbol{x} \in \mathbb{R}^N / \quad x_k \in]a_k, b_k[\,, \quad k = 1..N\}, \tag{2}$$

where $-\infty \leq a_k < b_k \leq \infty$ are known, $1 \leq k \leq N$.

Possible answers are given with set theoretic estimation and projection onto convex sets algorithms[1] (POCS). Although good reconstructions can be obtained using such approaches, they are often computationally expensive and do not lead to a unique and well-defined reconstruction. We do think that the importance of that drawback should not be overestimated but we will also present examples where the regular behavior of the reconstruction is used with benefits.

A different approach relies on the Bayesian setting. In this framework, the lack of an exact knowledge on a quantity is accounted for by defining a probability distribution over its possible values. Then the aim of an experiment is to provide a probability distribution for the quantities of interest to the observer. Roughly speaking, the more "sharp" is this so-called *a posteriori* distribution, the more information we have on the object.

Knowing how the experimental device behaves under a given solicitation allows the *direct* distribution $p(\boldsymbol{y}|\boldsymbol{x})$ to be defined. Then, the *a posteriori* distribution is given by the Bayes rule

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x})\, p(\boldsymbol{y}|\boldsymbol{x}) \tag{3}$$

which requires that the *prior* distribution $p(\boldsymbol{x})$ for $\boldsymbol{x}$ be also specified. The latter distribution summarizes what is known of the object before the experiment is performed. In a strict Bayesian sense, Eq. (3) gives the solution to the inverse problem since it gathers all information on $\boldsymbol{x}$. However, when the object consists of a large number of independent parameters, the study of its *a posteriori* distribution is cumbersome and often intractable. A single *a posteriori* estimation of the object is prefered, such as the Maximum of the *A Posteriori* distribution (MAP estimate). Taking the logarithm of Eq. (3) the MAP estimation reduces to the optimization of

$$\mathcal{J}(\boldsymbol{x}) = \log p(\boldsymbol{y} \mid \boldsymbol{x}) + \log p(\boldsymbol{x}).$$

Bayesian estimation is a satisfactory framework for reconstruction problems[4], yet in a situation where no *a priori* probabilistic model has "naturally" emerged or has been empirically found useful, the *ab initio* choice of a good *a priori* distribution $p(\boldsymbol{x})$ is a difficult task for which there is no general answer[7]. It is precisely the case when the only *a priori* knowledge on the object is a convex constraint such as Eq. (2). Choice of a prior is then guided by *ad hoc* considerations, among which the ability to easily compute the estimates is very important. It explains the success of gaussian models, which lead to quadratic regularization and linear (with respect to the data) estimates. Unfortunately these linear estimates usually cannot be guaranteed to satisfy Eq. (2).

2. Basis of a new approach and an early example

Deliberately leaving the Bayesian framework, we reformulate the problem as: "how to derive functionals $\mathcal{F}$ and $\mathcal{G}$ such that optimization of

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{x}) + \alpha \mathcal{G}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}). \tag{4}$$

yields a satisfactory reconstruction procedure ?" This question is made more precise with the following requirements or *desiderata* on the criterion $\mathcal{J}$:

A1 It should impose an exact fit to the data in the noiseless case, or a good fit to them while taking the noise statistics into account;

A2 It should impose the solution to be a member of $\mathcal{C}$;

A3 The reconstructed object should be uniquely defined as a regular function of the data, in order to ensure the uniqueness of the solution and to allow the study of its stability with respect to noise;

A4 When a *prior* guess $\boldsymbol{m}$ of the object is available (it can originate from a previous experiment for instance), the reconstruction process should have some properties of a projection of $\boldsymbol{m}$ onto the subset of all solutions that are consistent with the data and the constraints.

Note that A4 implies the following "natural" property: if the data are uninformative concerning the object under consideration then the reconstruction process should give back the prior guess $\boldsymbol{m}$ as a result. More details about the precise meaning of the fourth requirement can be found in reference[10].

Clearly, these few *desiderata* are not simultaneously satisfied by quadratic regularization or POCS methods. In contrast, we present now an early example of a regularized procedure, the so-called *maximum entropy* reconstruction of positive objects[6], which is in agreement with the *desiderata* in the case when $\mathcal{C} = \mathbb{R}_+^N$. It relies on the following criterion

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \alpha \sum_{i=1}^{N} \{x_i \log \frac{x_i}{m_i} - x_i + m_i\}, \tag{5}$$

where $\boldsymbol{m} = [m_1, m_2, \ldots, m_N]$ is a prior guess arising from previous measurements or chosen as a flat object. As far as the positivity constraint is concerned, criteria

like Eq. (5), built upon logarithmic expressions, ensure positivity and are therefore said to be "positivity free"; an other well-known example is the "$\log(x)$" or Burg entropy used in spectral analysis. Entropic regularization has been successfully used in several applied problems. Because of some rather acrimonious discussions in the literature, we emphasize that, in our view, maximum entropy reconstruction is not the ultimate answer to all positive inverse problems. We are simply interested in some of its properties (they are summarized in our four *desiderata*). Indeed, when dealing with arbitrary convex sets $\mathcal{C}$, it would be advantageous to exhibit a well behaved criterion just as Eq. (5). The presentation of an original constructive approach to obtain such criteria is the subject of the next section.

3. The Maximum Entropy on the Mean Method

The foundations of the Maximum Entropy on the Mean Method originate from the work of J. Navaza[11], and some theoretical aspects of the method were further studied by F. Gamboa and D. Dacunha-Castelle[2]. We have also studied it with a special attention to its potential applications in signal and image reconstruction and restoration[9]. For the sake of simplicity, this paragraph addresses the *noiseless* problem. Discussion of how to account for noise will take place in §6.

Much emphasis must be put on our only *a priori* information: the convex constraint of Eq. (2). The MEMM construction thus begins with the specification of the set $\mathcal{C}$ and a reference measure $\mathrm{d}\mu(\boldsymbol{x})$ over it.

Suppose that the actual observations $\boldsymbol{y}$ are the mean of a process $\boldsymbol{x}$ under a probability distribution P defined on $\mathcal{C}$ (this idea comes from statistical physics where observations are average values or macrostates). The set $\mathcal{C}$ being convex, the mean $\mathrm{E}_P\{\boldsymbol{x}\}$ under P is in $\mathcal{C}$ and hence the convex constraint is automatically fulfilled by $\mathrm{E}_P\{\boldsymbol{x}\}$.

3.1. Additional information principle

Since the constraint given by Eq. (2) does not lead to a unique distribution P, we have to invoke some additional information principle. For this purpose, we introduce the μ-entropy $K(P,\mu)$, or Kullback-Leibler (K-L) information[8]. This information is defined for a reference measure μ and a probability measure P by

$$\mathcal{K}(P,\mu) = \int \log \frac{\mathrm{d}P}{\mathrm{d}\mu}\,\mathrm{d}P \tag{6}$$

if P is absolutely continuous with respect to μ ($P \ll \mu$) and $\mathcal{K}(P,\mu) = +\infty$ otherwise.

We shall select the distribution P as the minimizer of the μ-entropy submitted to the constraints "on the mean" $\boldsymbol{A}\,\mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{y}$. In other words, P is the nearest distribution to the reference measure μ in the set of distributions such that $\boldsymbol{A}\,\mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{y}$, with respect to the K-L divergence. The maximum entropy on the mean problem then states as follows:

$$\text{MEMM problem} \qquad \begin{cases} \hat{P} = \arg\min\limits_{P} \displaystyle\int \log \frac{\mathrm{d}P}{\mathrm{d}\mu}(\boldsymbol{x})\mathrm{d}P(\boldsymbol{x}) \\ \text{such that } \boldsymbol{y} = \boldsymbol{A} \displaystyle\int \boldsymbol{x}\mathrm{d}P(\boldsymbol{x}) \end{cases}$$

It is well known that the solution, if it exists, belongs to in the exponential family

$$\mathrm{d}P_{\boldsymbol{s}}(\boldsymbol{x}) = \exp\left\{\boldsymbol{s}^{\mathrm{t}}\boldsymbol{x} - \log Z(\boldsymbol{s})\right\} \mathrm{d}\mu(\boldsymbol{x}), \tag{7}$$

and, more precisely, that its natural parameter is of the form $\boldsymbol{s} = \boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda}$. In Eq. (7) $\log Z$ is the log-partition function or the log-Laplace transform of the measure $\mathrm{d}\mu(\boldsymbol{x})$; for reasons that will become clear later, this function will be noted $\mathcal{F}^*$ in the sequel.

3.2. The dual problem

Using results of duality theory, there is an equality between the optimum value of the previous problem and the optimum value of its dual counterpart (dual attainment):

$$\operatorname*{Inf}_{P\in\mathcal{P}_y} \mathcal{K}(P,\mu) = \operatorname*{Sup}_{\boldsymbol{\lambda}\in\mathcal{D}_{\boldsymbol{\lambda}}} \left\{\boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda})\right\}, \tag{8}$$

where $\mathcal{P}_y = \{P : \boldsymbol{A}\,\mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{y}\}$ is the set of normalized distributions which satisfy the linear constraint on the mean, and $\mathcal{D}_{\boldsymbol{\lambda}}$ is the set $\{\boldsymbol{\lambda} \in \mathbb{R}^M : Z(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}) < \infty\}$, which is often the whole $\mathbb{R}^M$, in which case the dual problem is unconstrained.

Once the dual problem on the right side of Eq. (8) is solved, yielding an optimum value $\hat{\boldsymbol{\lambda}}$, one has the expression of the density $\hat{P} = P_{\boldsymbol{A}^{\mathrm{t}}\hat{\boldsymbol{\lambda}}}$ and can calculate the reconstructed object $\hat{\boldsymbol{x}}$ by computing (numerically) the expectation $\mathrm{E}_{\hat{P}}\{\boldsymbol{X}\}$. But this is not the more efficient way to compute the solution. Indeed, inside the exponential family (7) there is a one-to-one mapping between the natural parameter $\boldsymbol{s}$ and the mean of the associated distribution $\overline{\boldsymbol{x}}(\boldsymbol{s})$:

$$\overline{\boldsymbol{x}}(\boldsymbol{s}) = \frac{\mathrm{d}\mathcal{F}^*}{\mathrm{d}\boldsymbol{s}}(\boldsymbol{s})\,. \tag{9}$$

Therefore, the solution $\hat{\boldsymbol{x}}$ is simply obtained by calculating (9) at the optimal point $\boldsymbol{A}^{\mathrm{t}}\hat{\boldsymbol{\lambda}}$. We can now review the different steps for a practical implementation of the method. First, we have to choose a convex domain, $\mathcal{C}$, reflecting our knowledge about the domain where the solution has to be found, and a reference measure μ on this domain. Second, the log-partition function is calculated (analytically) together with the primal-dual relation of Eq. (9). Then the maximization of the dual criterion

$$D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}) \tag{10}$$

has to be (numerically) achieved. We emphasized that the dual criterion is by construction a strictly concave functional. Efficient methods of numerical optimization, such as gradient, conjugate gradient, or second order methods (Gauss-Newton) can be used to compute the solution. They will use the gradient of D which is easily calculated to be just $\boldsymbol{y} - \boldsymbol{A}\overline{\boldsymbol{x}}(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda})$. During the algorithm the primal-dual relation is used to compute the current reconstruction from the dual vector $\boldsymbol{\lambda}$.

3.3. Yet another primal problem

The previous development was done in the space of the dual parameters $\boldsymbol{\lambda}$. The purpose of this paragraph is to come back to the natural "object space". We will exhibit a new primal criterion, which we will call an entropy. This function, not surprisingly, is intimately related with the previous dual function and the K-L

information. Finally we will be able to derive, as particular cases of the MEMM procedure, many of the well known regularizing functionals.

For each $\boldsymbol{x} \in \mathcal{F}$, consider the MEMM problem when the constraint is $\mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{x}$. We define $\mathcal{F}(\boldsymbol{x})$ to be the optimum value of the K-L information for this problem

$$\mathcal{F}(\boldsymbol{x}) = \inf_{P \in \mathcal{P}_{\boldsymbol{x}}} \mathcal{K}(P, \mu),$$

where $\mathcal{P}_{\boldsymbol{x}} = \{P : \mathrm{E}_P\{\boldsymbol{X}\} = \boldsymbol{x}\}$.

As already seen, at the optimum, we have by dual attainment

$$\mathcal{F}(\boldsymbol{x}) = \sup_{\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\lambda}}} \left\{ \boldsymbol{\lambda}^{\mathrm{t}} \boldsymbol{x} - \mathcal{F}^*(\boldsymbol{\lambda}) \right\}, \tag{11}$$

The latter equation means that $\mathcal{F}$ is the conjugate convex of $\mathcal{F}^*$ and, as $\mathcal{F}^*$ is the log-Laplace transform of μ, the *Cramér transform* of μ. Such transforms appear in various fields of statistics and in particular in the Large Deviations theory, which has important connections with the MEMM[10]. Properties of Cramér transforms are listed below[5]:

- $\mathcal{F}$ is continuously differentiable and strictly convex on $\mathcal{C}$,
- $\mathcal{F}(\boldsymbol{x}) = +\infty$ for $\boldsymbol{x} \notin \mathcal{C}$ and its derivative is infinite on the boundary of $\mathcal{C}$,
- $\mathcal{F}(\boldsymbol{x}) \geq 0$ with equality for $\boldsymbol{x} = \boldsymbol{m}$, the mean value under the reference measure μ.

Our original MEMM problem can now be handled in a different way. If P is a candidate distribution with mean $\boldsymbol{x}$, its K-L information with respect μ is greater or equal to $\mathcal{F}(\boldsymbol{x})$. Moreover, this lower bound can be decreased by searching a vector $\hat{\boldsymbol{x}}$ minimizing $\mathcal{F}$ over the set $\mathcal{C}_{\boldsymbol{y}} = \{\boldsymbol{x} : \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{y}\}$. Then the MEMM problem is reformulated as

$$\inf_{\boldsymbol{x} \in \mathcal{C}_{\boldsymbol{y}}} \{ \inf_{P \in \mathcal{P}_{\boldsymbol{x}}} \mathcal{K}(P, \mu) \}.$$

If we consider the reconstruction problem in the object space, we only need to solve

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \in \mathcal{C}_{\boldsymbol{y}}} \mathcal{F}(\boldsymbol{x}). \tag{12}$$

Note that this problem has the same dual problem than that of Eq. (10). In fact, we have exhibited another primal problem associated to Eq. (10), directly in the object space $\mathbb{R}^N$. Its solution $\hat{\boldsymbol{x}}$ is the mean of the optimal distribution in the MEMM problem, and a solution to our reconstruction problem. This swap between primal problems is referred to as a "contraction principle" in statistical physics[5]. From this point of view, functional $\mathcal{F}$ appears as a level 1 entropy, therefore we will simply call it entropy in the following.

Properties of the Cramér transform are useful for reconstruction purposes, when holding the entropy $\mathcal{F}$ as the objective function, as in Eq. (12). Strict convexity enables a simple implementation and guarantees the uniqueness of the reconstruction. The second property shows that any descent method will provide a solution in $\mathcal{C}$, even if the constraint $\boldsymbol{x} \in \mathcal{C}$ is not specified in the algorithm; this

property, the "$\mathcal{C}$-free property", is here an analog of the "positivity free" property observed in maximum entropy reconstruction (see above). The last property shows that $\mathcal{F}$ may be considered as a discrepancy measure between $\boldsymbol{x}$ and $\boldsymbol{m}$. In the sequel, we give some examples illustrating the different points developed above.

4. A few examples of MEMM criteria

4.1. Gaussian reference

Our first example consists in a problem where no constraint is known on the object, so that $\mathcal{C} = \mathbb{R}^n$. We choose the Gaussian measure $\mathcal{N}(\boldsymbol{m}, \boldsymbol{R}_x)$ as our reference measure μ on $\mathcal{C}$. A simple calculus then leads to the Cramér transform

$$\mathcal{F}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{m})^{\mathrm{t}} \boldsymbol{R}_x^{-1} (\boldsymbol{x} - \boldsymbol{m}), \tag{13}$$

which is recognized as a quadratic regularizing term, already mentioned as being linked to a Gaussian prior distribution with expectation $\boldsymbol{m}$ and covariance matrix $\boldsymbol{R}_x$.

4.2. The positive case

• Poisson reference and the "Shannon entropy"

Let now $\mathcal{C}$ be $]0, +\infty[$, and the reference distribution be a Poisson law, with expectation $\boldsymbol{m}$. As usual, without any information regarding correlation between adjacent pixels, the distribution is supposed separable. Such a prior may correspond to the modeling of the fall of quanta of energy, following a Poisson process, in such a way that the expectation at a site j is m_j. This modeling may be encountered in astronomy (the speckle-images of optical interferometry) for instance. The reference measure is then

$$\mu(\boldsymbol{x}) = \prod_{j=1}^{N} \mu(x_j) = \prod_{j=1}^{N} \frac{m_j^{x_j}}{x_j!} \exp(-m_j).$$

Observations are again a linear transform $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ of an unknown object $\boldsymbol{x}$, and we select as a solution the mean under the nearest distribution to μ satisfying the constraint for that mean, the distance being measured by the Kullback-Leibler distance. The entropy functional $\mathcal{F}$ is the Cramér transform of μ, and works out to be

$$\mathcal{F}(\boldsymbol{x}) = \sum_{j=1}^{N} \left[\frac{x_j}{m_j} \log \left(\frac{x_j}{m_j} \right) + m_j - x_j \right],$$

which is the generalized version of the Shannon entropy (the corrective term $m_j - x_j$ ensures the positivity of $\mathcal{F}$ when either $\boldsymbol{x}$ or $\boldsymbol{m}$, or both, are not normalized to unity).

• Gamma reference and Itakura-Saïto discrepancy measure

The problem takes place in spectrum analysis. We review here the presentation of reference[12], which happens to be exactly a MEMM approach to a well known criterion: the Itakura-Saïto discrepancy measure.

Let the data $\boldsymbol{y}$ be a vector of autocorrelation samples. We consider here only the finite-dimensional problem, *i.e.* estimation of the power spectra over a list of k frequencies. The relation between the data $\boldsymbol{y}$ and the (discretized) spectrum $\boldsymbol{x}$ is the Fourier transform $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is a $M \times N$ Fourier matrix, with M the number of known correlation samples and N the number of wanted spectrum samples. The periodogram having asymptotically a χ^2 distribution with two degrees of freedom, the corresponding reference measure μ over the possible spectra is an exponential law with mean, *i.e. prior spectrum* $\boldsymbol{m}$. Using the Cramér transform definition, one easily obtains the entropy

$$\mathcal{F}(\boldsymbol{x}) = \sum_{j=1}^{N} \frac{x_j}{m_j} - \log\left(\frac{x_j}{m_j}\right) - 1, \tag{14}$$

which is the Itakura-Saïto distortion between $\boldsymbol{s}$ and $\boldsymbol{m}$. Observe that $\boldsymbol{m}$, which is the mean under μ, is also the minimum of the Itakura-Saïto distortion without constraint, and is therefore the prior guess. With $\boldsymbol{m} = \boldsymbol{1}$, we measure a distance to a flat spectrum, and find out the so-called "$\log(x)$", or Burg entropy.

4.3. The bounded case

We consider here the case when $\mathcal{C}$ has the general form of Eq. 2. Such constraints may be useful in many applied problems where the object is *a priori* known to lie between two bounds (tomography, filter design, crystallography).

Several reference measures can be used on the convex $\mathcal{C}$. A natural idea is indeed to use a product of uniform measures over each interval $]a_j, b_j[$:

$$\mathrm{d}\mu(\boldsymbol{x}) = \bigotimes_{j=1}^{N} \frac{1}{b_j - a_j} \mathbf{1}_{]a_j, b_j[}(x_j)\, \mathrm{d}x_j\,.$$

The calculus of the Cramér transform leads to implicit equations, therefore we have no analytic expression for $\mathcal{F}$. Nevertheless the primal-dual relation can be computed

$$x_j = -\frac{1}{s_j} + \frac{b_j e^{b_j s_j} - a_j e^{a_j s_j}}{e^{b_j s_j} - e^{a_j s_j}} \quad \text{with } s_j = [\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}]_j, \quad 1 \leq j \leq N$$

and the convex problem

$$\begin{cases} \operatorname{Inf}_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x}) \\ \\ \text{subject to } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \end{cases}$$

where $\mathcal{F}$ is *not explicit*, can still be solved using its dual formulation

$$D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}),$$

together with the aforementioned primal-dual relation. Other measures could be used in this case. The case of a Bernoulli measures product $\mathrm{d}\mu(\boldsymbol{x}) = \bigotimes_{j=1}^{N} \{\alpha_j \delta(x_j - a_j) + (1-\alpha_j)\delta(x_j - b_j)\}$ (where δ denotes the Dirac measure) is derived in a referenced work[10].

5. Illustrations

Two illustrations are given here. The first one concerns data from the Hubble Space Telescope, and the second one is a synthetic example in Fourier synthesis.

Fig. 1.a

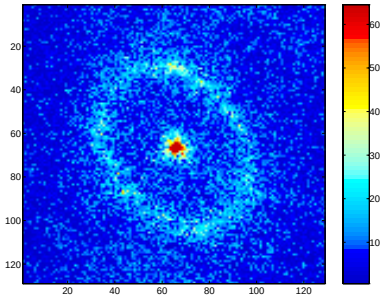

Fig. 1.b

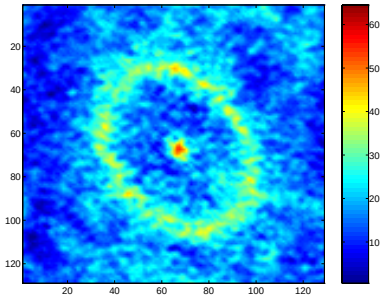

Fig. 1.c

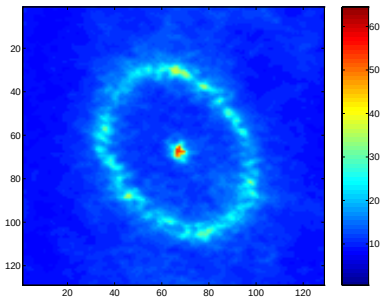

Figure 1

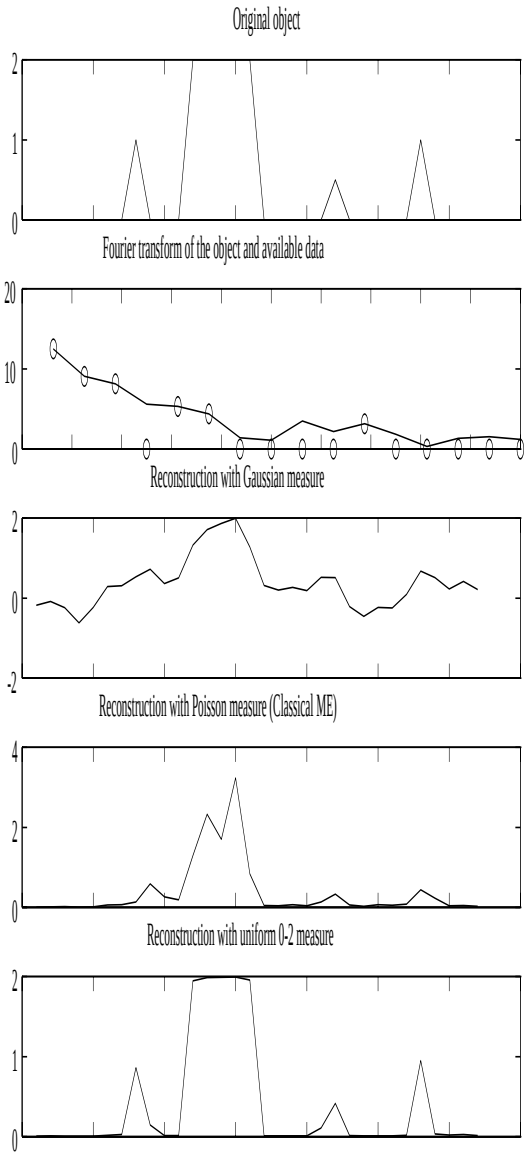

Figure 2

Figure 1 — An image FOC-f/96 (*Faint Object Camera*) of Supernova SN1987A which exploded in February 1987 is given in Fig. 1.a. Data are blurred by the large impulse response of the Hubble Space Telescope (HST) (before its correction in January 1994). Figs. 1.b and 1.c compare two restorations of this image; the first being a standard Maximum Entropy one, while the second is obtained with the MEMM. Due to their very large dynamics, the results of Figs. 1.b and 1.c are in logarithmic scale. This shows the improvement provided by the MEMM, especially concerning the background of the sky.

Figure 2 — This figure compares different reconstruction methods in a simple Fourier synthesis problem. The original object and the available data (o) are on the top. Then three reconstructions, corresponding to different reference measures in the MEMM scheme, and also to different constraint sets $\mathcal{C}$, are given. They show an improvement with the reduction of the "admissible set" of solutions.

6. Taking noise into account

So far MEMM criteria have been derived from the maximization of the μ-entropy submitted to an *exact constraint*. Any observation noise will ruin our exact constraint, and as a consequence the two (primal-dual) formulations of the MEMM problem. The exact constraint was useful in interpreting observations as a linear transform of a mean, then enabling us to exhibit the discrepancy measure $\mathcal{F}$. Because of the good properties of $\mathcal{F}$, we will keep on considering the unknown object $\boldsymbol{x}$ as a mean, in order to use its entropy $\mathcal{F}(\boldsymbol{x})$, but we will have to modify the procedure. In the sequel, we first introduce the noise using a χ^2 constraint, then we turn towards a modification of the MEMM setting to explicitely account for the noise.

6.1. The χ^2 constraint

A classical way to account for noise is to construct a confidence region about the expected value of some statistic. For gaussian noise, one usually uses the χ^2 constraint $||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2 \leq \rho$, where ρ is some constant. Then the problem becomes the minimization of $\mathcal{F}$ submitted to the χ^2 constraint. There always exists a positive parameter α (in fact it is a Lagrange parameter corresponding to the χ^2 constraint) such that the previous problem and the penalized problem

$$\inf_{\boldsymbol{x}} \{\mathcal{F}(\boldsymbol{x}) + \alpha||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||^2\} \tag{15}$$

have the same solution. Since we may not have an analytic expression for $\mathcal{F}$, while we always have an expression of its conjugate $\mathcal{F}^*$, our goal is to exhibit a problem equivalent to Eq. (15) expressed in terms of $\mathcal{F}^*$. In order to achieve that, we need to transform Eq. (15) into a constrained problem. The idea, according to the reference[3], is to introduce a new set of parameters, namely $\boldsymbol{\xi}$, and to replace the optimization of Eq. (15) by the equivalent

$$\begin{cases} \inf\limits_{\boldsymbol{x},\boldsymbol{\xi}} \{\mathcal{F}(\boldsymbol{x}) + \alpha||\boldsymbol{\xi}||^2\}, \\ \boldsymbol{\xi} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}. \end{cases} \tag{16}$$

The Lagrangian of Eq. (16) is

$$\tilde{L}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = L(\boldsymbol{x}, \boldsymbol{\lambda}) + \alpha||\boldsymbol{\xi}||^2 + \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{\xi},$$

where L is the Lagrangian of the noiseless problem. This Lagrangian can be minimized separately with respect to $\boldsymbol{x}$ and $\boldsymbol{\xi}$, leading to the new dual function $\tilde{D}$:

$$\tilde{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - K^*(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}) - \frac{1}{2\alpha}||\boldsymbol{\lambda}||^2 = D(\boldsymbol{\lambda}) - \frac{1}{2\alpha}||\boldsymbol{\lambda}||^2,$$

with $\boldsymbol{\xi} = -\boldsymbol{\lambda}/(2\alpha)$. Thus, the only modification given by the addition of a penalization in the direct problem is the addition of a regularizing term in the dual function. The primal-dual relation remains the same and the reconstruction is again ensured to belong to the specified convex set $\mathcal{C}$.

6.2. Accounting for general noise statistic within the MEMM *procedure*

Thanks to a specific entropy function, more complicated penalizations than Eq. (15) can be performed in order to account for non-gaussian noises. Such entropies can be derived directly in the same MEMM axiomatic approach as in the noiseless case. To this end, we only need to introduce an *extended object* $\tilde{\boldsymbol{x}} = [\boldsymbol{x},\ \boldsymbol{n}]$, and consider the relation $\boldsymbol{y} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}$, with $\tilde{\boldsymbol{A}} = [\boldsymbol{A},\ \boldsymbol{1}]$. The vector $\tilde{\boldsymbol{x}}$ evolves in the convex $\tilde{\mathcal{C}}$ of $\mathbb{R}^{N+M}$, which separates on a product of the usual $\mathcal{C}$ and of $\mathcal{B}$, $\tilde{\mathcal{C}} = \mathcal{C}\times\mathcal{B}$, where $\mathcal{B}$ is the convex hull of the state space of the noise vector $\boldsymbol{n}$.

We then use a reference measure ν over the noise set. For instance, in the case of a Gaussian noise we take $\mathcal{B} = \mathbb{R}^M$ and a centered gaussian law with covariance matrix $\boldsymbol{R}_\nu$ as ν. With a Poisson noise we take $\mathcal{B} = \mathbb{R}_+^M$ and a Poisson reference measure ν.

Now we can define a new entropy functional by using a reference measure $\tilde{\mu}$ on $\tilde{\mathcal{C}}$. If ν is the distribution of the noise, μ our object reference measure on $\mathcal{C}$ and if we assume that the object and noise are independent, we obtain $\tilde{\mu} = \mu \otimes \nu$. The entropy function we looked for is then the Cramér transform of $\tilde{\mu}$ which is simply

$$\mathcal{F}_{\tilde{\mu}}(\tilde{\boldsymbol{x}}) = \mathcal{F}_\mu(\boldsymbol{x}) + \mathcal{F}_\nu(\boldsymbol{n}).$$

Estimation of the extended object is conducted through a constrained minimization of $\mathcal{F}_{\tilde{\mu}}(\tilde{\boldsymbol{x}})$, the constraint being $\boldsymbol{y} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$. Therefore it reduces to the unconstrained minimization of the compound criterion

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_{\tilde{\mu}}([\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}]^{\mathrm{t}}) = \mathcal{F}_\mu(\boldsymbol{x}) + \mathcal{F}_\nu(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}). \tag{17}$$

A dual approach is again useful, in particular if $\mathcal{F}_\mu$ or $\mathcal{F}_\nu$ or both are not explicit. It is easy to show that the dual criterion is

$$\tilde{\mathcal{D}}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - \mathcal{F}_\mu^*(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}) - \mathcal{F}_\nu^*(\boldsymbol{\lambda}).$$

Having solved the dual problem we come back to the primal solution by the primal-dual relation which is, thanks to the separability of the log-Laplace transform of $\tilde{\mu}$, the same as in the noiseless case:

$$\overline{\boldsymbol{x}}(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}) = \frac{\mathrm{d}\mathcal{F}_\mu^*}{\mathrm{d}\boldsymbol{s}}(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}).$$

We are then able to account for specific noise distributions, without loss in the nice properties of our criteria: the global criterion of Eq. (17) is always convex, and the convex constraint is automatically satisfied. Concerning the case of a

Gaussian noise, it can easily be checked using result of Eq. (13), that a gaussian reference measure for the noise term leads to the problem of Eq. (15), obtained by statistical considerations.

It is always possible to modify our reference measures to balance the two terms of the global criterion Eq. (17) which should therefore be written as

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_\mu(\boldsymbol{x}) + \alpha \mathcal{F}_\nu(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}),$$

where α is a regularization parameter. The Maximum Entropy on the Mean procedure enables us to find the generic form of regularized criteria, and to solve the problem even if primal criteria $\mathcal{F}_\mu$ and $\mathcal{F}_\nu$ have no analytical expression.

Such an approach provides a new general framework for the interpretation and derivation of these criteria. Many other criteria as those presented in §4 have been derived[10]. In particular, reference measures defined as mixture of distributions (Gaussian, Gamma) have been successfully used for the reconstruction of blurred and noisy sparse spike trains. Poissonized sums of random variables also lead to interesting regularized procedure in connection with the general class of Bregman divergences[10]. Work is also in progress concerning the *quantification* of the quality of MEMM estimates, the links with the Bayesian approach, especially with correlated *a priori* models such as Gibbs random fields.

References

1. P. L. Combettes, *Proceedings of the IEEE* **81** (1993) 182.
2. D. Dacunha-Castelle and F. Gamboa, *Ann. Inst. Henri Poincaré* **26** (1990) 567.
3. A. Decarreau, D. Hilhorst, C. Lemaréchal and J. Navaza, *SIAM J. Optimization* **2** (1992) 173.
4. G. Demoment, *IEEE Trans. on Acoust., Speech, and Signal Proc.* **37** (1989) 2024.
5. R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics* (Springer-Verlag, New York, 1985).
6. S. F. Gull and G. J. Daniell, *Nature* **272** (1978) 686.
7. Kass and Wasserman, unplublished manuscript (1994).
8. S. Kullback, *Information theory and statistics* (Wiley, New York, 1959).
9. G. Le Besnerais, *PhD thesis* (University of Paris, 1993).
10. G. Le Besnerais, J.-F. Bercher and G. Demoment, submitted to *IEEE Trans. on Information Theory*, (1994).
11. J. Navaza, *Acta Cryst.* **A-42** (1986) 212.
12. J. E. Shore, *IEEE Trans. Acoust., Speech and Signal Proc.* **29** (1981) 230.