A New Look at Entropy for Solving Linear Inverse Problems

Guy Le Besnerais, Jean-François Bercher, and Guy Demoment

***Abstract*—Entropy-based methods are widely used for solving inverse problems, particularly when the solution is known to be positive. Here, we address linear ill-posed and noisy inverse problems of the form $z = Ax + n$ with a general convex constraint $x \in \mathcal{X}$, where $\mathcal{X}$ is a convex set. Although projective methods are well adapted to this context, we study alternative methods which rely highly on some "information-based" criteria. Our goal is to clarify the role played by entropy in this field, and to present a new point of view on entropy, using general tools and results coming from convex analysis. We present then a new and broad scheme for entropic-based inversion of linear-noisy inverse problems. This scheme was introduced by Navaza in 1985 in connection with a physical modeling for crystallographic applications, and further studied by Dacunha-Castelle and Gamboa. Important features of this paper are: i) a unified presentation of many well-known reconstruction criteria, ii) proposal of new criteria for reconstruction under various prior knowledge and with various noise statistics, iii) a description of practical inversion of data using the aforementioned criteria, and iv) a presentation of some reconstruction results.**

***Index Terms*— Duality, entropy, inverse problems, Kullback–Leibler information, regularization.**

I. Introduction

INVERSE problems appear in most experimental data processing problems. In such situations, one has to derive an estimate $\hat{\boldsymbol{x}}$ of some physical quantity of interest (the object) from indirect and noisy observations $\boldsymbol{z}$. In this paper, we focus on the "linear-noisy" observation model

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

where the transfer matrix $\boldsymbol{A}$ and some statistical characteristics of the noise $\boldsymbol{n}$ are known.

In many practical problems, the object is often known *a priori* to satisfy a convex constraint such as

$$\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^K. \tag{2}$$

Examples of such problems are encountered when imaging intensity distributions (for instance, in astronomy or in spectral analysis) which are known, on physical grounds, to be positive. In other specific problems, such as crystallography and tomography, lower and upper bounds on the intensity are known, see [1]. These constraints may generally be described by (2), with

$$\mathcal{X} = \{\boldsymbol{x}\colon a_j \le x_j \le b_j,\ 1 \le j \le K\} \quad \text{where } -\infty \le a_j < b_j \le \infty \text{ for each } j. \tag{3}$$

Within many finite-dimensional inverse problems, $\boldsymbol{A}$ is a nonregular ill-conditioned matrix. Therefore, a simple generalized inverse gives unsatisfactory results, because of the dramatic amplification of any observation noise. Many studies have been devoted to the task of designing reconstruction procedures yielding stable reconstructions satisfying (2) or (3). In this paper, we use entropy to build a general framework for inversion under convex constraints, including well-known reconstruction techniques as particular cases.

A. The Penalization Approach

A popular approach to linear inverse problems resolution is the minimization of penalized cost functions of the form

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{x}, \boldsymbol{z}) + \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{m}) \tag{4}$$

that corresponds to the general scheme of balancing between trust to data and fidelity to some priors, where the penalizing term $\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{m})$ can, for instance, force the solution to belong to the constraints set $\mathcal{X}$, and may depend on a prior object $\boldsymbol{m}$. These criteria may originate from practical considerations on the desired properties of the solution (for instance, it can be designed in order to smooth the solution) or can derive from a preliminary modelization step, for instance a probabilistic modelization in Bayesian estimation. Important properties of the reconstruction process are linked to the properties of $\mathcal{F}_{\boldsymbol{n}}$ and $\mathcal{F}_{\boldsymbol{x}}$ in the general criterion $\mathcal{J}$.

Several authors have presented axiomatic approaches of inversion, that lead to compound criteria, whose properties (as functions of $\boldsymbol{x}$, $\boldsymbol{z}$, and $\boldsymbol{m}$) translate the chosen axioms, see, for example, [2]–[6]. In this case, the resulting criteria have by construction good properties, but these constructions are limited, to our knowledge, to the convex constraint $\boldsymbol{x} \in \mathcal{X}$, with

$$\mathcal{X} = \mathbb{R}^K \qquad \mathcal{X} = (\mathbb{R}^+)^K$$
$$\text{or } \mathcal{X} = \left\{\boldsymbol{x} \in (\mathbb{R}^+)^K\colon \sum_{k=1}^{K} x_k = 1\right\}. \tag{5}$$

Manuscript received September 12, 1995; revised May 26, 1998. This work was performed while the authors were with the Laboratoire des Signaux et Systèmes, Plateau de Moulon, 91192 Gif-sur-Yvette, France. This paper is dedicated in memory of Edwin T. Jaynes, indefatigable advocate of Entropy and Bayesian approaches.

G. Le Besnerais is with ONERA, DTIM/TI, 92322 Châtillon Cedex, France (e-mail: G.Lebesnerais@onera.fr).

J.-F. Bercher is with the Département Signaux et Télécoms, ESIEE, 93162 Noisy-le-Grand Cedex, France (e-mail: bercher@esiee.fr).

G. Demoment is with the Laboratoire des Signaux et Systèmes, Plateau de Moulon, 91192 Gif-sur-Yvette, France (e-mail: Demoment@lss.supelec.fr).

Communicated by A. Hero, Associate Editor for Signal Processing.

Publisher Item Identifier S 0018-9448(99)04734-3.

B. Proposed Approach

In the sequel, we present an entropy-based framework to design reconstruction criteria of the form (4). It makes a trade-off between axiomatic constructions and general penalized approaches, in the sense that criteria derived in this framework share interesting properties and that some *degrees of freedom* are left to the user in order to design particular criteria adapted to a particular reconstruction problem.

The resulting "entropic inversion framework" was introduced in 1985 by Navaza [7], [8] while dealing with a particular reconstruction problem which occurs in crystallography and which exhibits constraints like (3). It has been further studied since 1989 by mathematicians, see [9]–[13], with emphasis on the extension over spaces of functions or measures. Other studies emphasize applications to various finite-dimensional linear inverse problems [14]–[18]. Similar results, including account for observation noise and correlation issues, have been introduced independently in [19].

In this paper, we aim to give a systematic presentation of the entropic framework for inversion. From a practical point of view, the methodology will finally yield the regularized criterion (4), where prior knowledge will be encoded into the expressions and properties of $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{n}}$. The resulting reconstruction process reduces to an unconstrained minimization of $\mathcal{J}$. Among their properties, $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{n}}$ will be strictly convex by construction, therefore, the minimization of the global criterion $\mathcal{J}$ is greatly simplified.

Section II is devoted to the derivation of this entropic framework: we state the problem and useful mathematical results, give the general form of the resulting entropic criteria, together with their important dual formulation. We then examine common properties of these criteria. In Section III, we present some particular applications of our framework, where it yields particular criteria either known or new. Section IV is then devoted to the application of the procedure to inverse problems. We also give hints for some of the remaining practical choices in order to use an entropic inversion method. Finally, we give some inversion examples.

II. Building Cost Functions using Entropy

In this section, we derive the general form of cost function (4), taking into account convex constraints on the sought object and the fact that data are a linear transformation of the original unknown object.

The observation noise will be explicitly accounted for: in fact, we will consider the linear noisy initial problem as a linear problem relating the data to the pair of unknown vectors $(\boldsymbol{x}, \boldsymbol{n})$

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n} = [\boldsymbol{A}, \boldsymbol{I}]\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{n}\end{bmatrix} \tag{6}$$

or

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{y},$$

with $\boldsymbol{H} = [\boldsymbol{A}, \boldsymbol{I}]$, and $\boldsymbol{y}^t = [\boldsymbol{x}^t \ \ \boldsymbol{n}^t]$, and our goal is to estimate both vectors $\boldsymbol{x}$ and $\boldsymbol{n}$ from data $\boldsymbol{z}$ while using prior information.

A. Introducing Entropy

The method presented here relies heavily on Kullback–Leibler information (also known as I-divergence), which is defined for some reference measure μ and probability measure P by

$$\mathcal{K}(P, \mu) = -H_\mu(P) = \int \log \frac{dP}{d\mu}\, dP \tag{7}$$

if P is absolutely continuous with respect to μ ($P \ll \mu$), and $\mathcal{K}(P, \mu) = +\infty$ otherwise. $H_\mu(P)$, the negative of the Kullback–Leibler information, is the μ-entropy of P. Entropy has been a subject of many studies since its introduction in engineering by Shannon (and also independently by Wiener) [20]. If μ is a probability measure, then $\mathcal{K}(P, \mu) \geq 0$ with equality if and only if $P = \mu$. $\mathcal{K}$ is used as a discrepancy measure in statistics, where it has been introduced and studied by Kullback and Leibler, see [5] and [21]–[24].

When dealing with inverse problems, entropy has been invoked for two purposes: first, it can be used as a regularizing term for reconstruction of positive objects (it is the "$x \log x$" entropic regularization method, see Section III-B); second, it is used for selecting probabilistic distributions. Actually, given some "testable informations," which are usually moment constraints, Jaynes' maximum entropy principle may be used to select a probability distribution [25]. For these two uses of entropy, see also papers in [26].

B. A Statistical Physics Model

Like many recent models used in image processing, the proposed approach can be illustrated by a statistical physics model. Let us consider a system consisting of a large number N of independent particles whose state $\{\boldsymbol{Y}_n\}_{1\leq n\leq N}$ may evolve within the state space $S_\mu \subset \mathbb{R}^K$. Each state of S_μ is more or less probable according to reference probability measure μ, which reflects some physical property (for instance, the nature of the particles) and whose mean is denoted by $\boldsymbol{m}$. Under well-known assumptions (see, for instance, [27]), the mean state of the particles $\overline{\boldsymbol{Y}}_N$ converges to $\boldsymbol{m}$, while the empirical law $\hat{P}_N$ converges to μ for large N. Therefore, μ and $\boldsymbol{m}$ are considered as two possible macroscopic descriptions (or macrostates) of the equilibrium of the system. In [28], $\boldsymbol{m}$ is called the level-1 equilibrium macrostate associated with the system, and μ its level-2 equilibrium macrostate.

Suppose now that a macroscopic linear observation of the following form is available:

$$\boldsymbol{z} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{H}\boldsymbol{Y}_n = \boldsymbol{H}\overline{\boldsymbol{Y}}_N. \tag{8}$$

The problem is now to take (8) into account to update the equilibrium macroscopic descriptions (or macrostates) of the system. Let us call $\mathcal{L}$ the set $\{\boldsymbol{y} \in \mathbb{R}^K : \boldsymbol{H}\boldsymbol{y} = \boldsymbol{z}\}$ and $\mathcal{P}$ the set of probability distributions over S_μ whose mean value belongs to $\mathcal{L}$. If $\boldsymbol{m} \in \mathcal{L}$ (or, equivalently, if $\mu \in \mathcal{P}$), the data (8) do not contradict the initial macroscopic description $(\boldsymbol{m}, \mu)$, which therefore remains unchanged. If $\boldsymbol{m} \notin \mathcal{L}$, new macrostates $(\hat{P} \in \mathcal{P}, \hat{\boldsymbol{y}} \in \mathcal{L})$ have to be found. The new level-2 description is given by the maximum μ-entropy distribution,

see the work of Jaynes [25]

$$\hat{P} = P^{\mathrm{ME}} = \underset{P\in\mathcal{P}}{\operatorname{argmax}}\, H_\mu(P) = \underset{P\in\mathcal{P}}{\operatorname{argmin}}\, \mathcal{K}(P,\mu) \quad (9)$$

(note that this expression supposes that the infimum of the right-hand side is attained in $\mathcal{P}$, see [29] for general results on this problem). This result can be related to Sanov's theorem, see for example [28], [30], and [31], with the "relaxed observation" set $\mathcal{L}_\epsilon = \{\boldsymbol{y} \in \mathbb{R}^K : \|\boldsymbol{z} - \boldsymbol{H}\boldsymbol{y}\| < \epsilon\}$ for some $\epsilon > 0$ and with $\mathcal{P}_\epsilon$ the set of probability distributions over S_μ whose mean value belongs to $\mathcal{L}_\epsilon$. We assume that $\boldsymbol{m} \notin \overline{\mathcal{L}}_\epsilon$ and emphasize the case of a finite state space S_μ in the sequel. In this case, the method of types [24], [32] gives the asymptotic behavior of the empirical distribution

$$P(\hat{P}_N = P) = \exp\{-N\mathcal{K}(P,\mu) + O(\log(N))\} \quad (10)$$

where P denotes a possible N-size sample empirical distribution. It leads to the probability of the set $\mathcal{P}_\epsilon$

$$\lim_{N\to\infty} \frac{1}{N} \log P(\{\hat{P}_N \in \mathcal{P}_\epsilon\}) = -\min_{P\in\mathcal{P}_c} \mathcal{K}(P,\mu) \quad (11)$$

assuming

$$\exists\, P \in \operatorname{int}\mathcal{P}_\epsilon : \mathcal{K}(P,\mu) < +\infty.$$

Equations (10) and (11) explain the choice of the maximum entropy distribution P^{ME} of (9) as the new level-2 description of the system [33]. This argument is classical in the literature on applications of the entropy and known as the *multiplicity argument*, see the papers of Jaynes [25]. Moreover, it can be linked to a physical model of image formation, see Section III-B.

Here we are more interested in the level-1 description, or level-1 macrostate $\hat{\boldsymbol{y}}$. Intuitively, it should be defined as the mean of P^{ME}

$$\hat{\boldsymbol{y}} = E_{P^{\mathrm{ME}}}[\boldsymbol{Y}].$$

An alternate way to define this description is to study directly the rate of the exponential decay of the probability of the events $\{\overline{\boldsymbol{Y}}_N \in \mathcal{L}_\epsilon\}$. Cramer's theorem [28], [30], [31] provides a "rate function" $\mathcal{F}$, under some assumptions on the reference measure μ. Then

$$\lim_{N\to\infty} \frac{1}{N} \log P(\{\overline{\boldsymbol{Y}}_N \in \mathcal{L}_\epsilon\}) = -\min_{\boldsymbol{y}\in\mathcal{L}_c} \mathcal{F}(\boldsymbol{y})$$

assuming that $\exists\, N_0$ such that $P(\{\hat{\boldsymbol{Y}}_N \in \mathcal{L}_\epsilon\}) > 0$ for $N > N_0$ and

$$\min_{\boldsymbol{y}\in\overline{\mathcal{L}}_c} \mathcal{F}(\boldsymbol{y}) = \min_{\boldsymbol{y}\in\operatorname{int}\mathcal{L}_c} \mathcal{F}(\boldsymbol{y}).$$

The new level-1 description of the system should thus be defined as

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}\in\mathcal{L}}{\operatorname{argmin}}\, \mathcal{F}(\boldsymbol{y}). \quad (12)$$

(At this stage, we suppose that the infimum at the right-hand side is attained.) The so-called *level-1 entropy* $\mathcal{F}$ is defined over S_μ: actually it is defined over the closed and convex hull of the state space S_μ.

Changing from the level-2 entropy $\mathcal{K}$ to the level-1 entropy $\mathcal{F}$ is called a *contraction principle* in large deviations theory. It is the foundation of the proposed framework for reconstruction criteria design. The resulting advantages can already be seen: being deduced from the "usual" entropy $\mathcal{K}$, the functions $\mathcal{F}$ will share many of its properties. At the same time, various choices of the reference measure will lead to different cost functions $\mathcal{F}$: this is the "degree of freedom" that was announced in Section I. In the next section, we will study further the links between reference measures μ and level-1 entropies $\mathcal{F}$, and between the two optimization problems (9) and (12).

C. Maximum Entropy on the Mean Description

1) Problem Statement: Let us recall that the problem statement is

$$P^{\mathrm{ME}} = \underset{P\in\mathcal{P}}{\operatorname{argmax}}\, H_\mu(P) = \underset{P\in\mathcal{P}}{\operatorname{argmin}}\, \mathcal{K}(P,\mu). \quad (13)$$

Having found this distribution, the level-1 description of the system is given by its mean value $\hat{\boldsymbol{y}} = E_{P^{\mathrm{ME}}}[Y]$. Problem (13) is called *maximum entropy on the mean problem* by Dacunha-Castelle [9].

There are many references concerning the existence of a solution for entropy or entropy-like minimization problems, cf. [29], [34], [35].

Along the line of the preceding section, we propose here another interpretation, by using the double stage formulation

$$\min_{P:\,\boldsymbol{z}=\boldsymbol{H}E_P[\boldsymbol{Y}]} \mathcal{K}(P,\mu) = \min_{\boldsymbol{y}:\boldsymbol{H}\boldsymbol{y}=\boldsymbol{z}} \left\{ \min_{P:\,E_P[\boldsymbol{Y}]=\boldsymbol{y}} \mathcal{K}(P,\mu) \right\}. \quad (14)$$

Equation (14) is only a formal guideline for the proposed approach. Actually, we will define in Proposition 1 a convex function $\mathcal{F}$ on $\mathbb{R}^k$ such that

$$\mathcal{F}(\boldsymbol{y}) \le \min_{P:\,E_P[\boldsymbol{Y}]=\boldsymbol{y}} \mathcal{K}(P,\mu), \qquad \forall\, \boldsymbol{y}$$

and study cases of equality. Then we will study the convex optimization problem

$$\min_{\boldsymbol{y}:\boldsymbol{H}\boldsymbol{y}=\boldsymbol{z}} \mathcal{F}(\boldsymbol{y})$$

(see Proposition 2). Finally, in Proposition 3, we will give sufficient conditions for existence of a vector $\hat{\boldsymbol{y}}$ such that

$$\begin{aligned}\min_{P:\boldsymbol{z}=\boldsymbol{H}E_P[\boldsymbol{Y}]} \mathcal{K}(P,\mu) &= \min_{\boldsymbol{y}:\boldsymbol{H}\boldsymbol{y}=\boldsymbol{z}} \mathcal{F}(\boldsymbol{y}) = \mathcal{F}(\hat{\boldsymbol{y}}) \\ &= \min_{P:E_P[\boldsymbol{Y}]=\hat{\boldsymbol{y}}} \mathcal{K}(P,\mu)\end{aligned}$$

and a duality result.

2) Preliminary Definitions: The *log-Laplace transform* of reference measure μ is

$$\mathcal{F}^*(\boldsymbol{s}) \triangleq \log \int \exp(\boldsymbol{s}^t\boldsymbol{u})\, d\mu(\boldsymbol{u}). \quad (15)$$

We consider $\mathcal{F}^*$ as an extended valued function, which means that it can take the value $+\infty$. The *effective domain* $D_{\mathcal{F}^*}$ (we will say domain in the sequel) of $\mathcal{F}^*$ is

$$D_{\mathcal{F}^*} \triangleq \{\boldsymbol{s} \in \mathbb{R}^K : \mathcal{F}^*(\boldsymbol{s}) < \infty\}.$$

$\mathcal{F}^*$ is convex and lower semicontinuous (l.s.c.), a property which is equivalent to closedness of the level sets $\{\boldsymbol{s}\colon \mathcal{F}^*(\boldsymbol{s}) \leq \alpha\}$ for every $\alpha \in \mathbb{R}$. We will suppose in the sequel that μ is *maximal*: its support S_μ is not contained in any proper subspace of $\mathbb{R}^K$. This property implies that $\mathcal{F}^*$ is strictly convex [28, p. 260]. Note that maximality of μ is not strictly necessary, it is used for the sake of simplicity.

Associated with the reference measure μ is the exponential family $\xi_\mu = \{P_{\boldsymbol{s}}\}$, $\boldsymbol{s} \in D_{\mathcal{F}^*}$. For each $\boldsymbol{s} \in D_{\mathcal{F}^*}$, $P_{\boldsymbol{s}}$ is the probability measure on $\mathbb{R}^K$ absolutely continuous with respect to μ, whose Radon–Nikodym derivative is

$$\frac{dP_{\boldsymbol{s}}}{d\mu} = \exp\bigl(\boldsymbol{s}^t\boldsymbol{u} - \mathcal{F}^*(\boldsymbol{s})\bigr).$$

Let us suppose that the domain of $\mathcal{F}^*$ has interior points: $\operatorname{int} D_{\mathcal{F}^*} \neq \emptyset$. Using properties of the Laplace transform of μ, it can be proved that $\mathcal{F}^*$ is differentiable at any interior point of its domain (see [28] and [36]), and $\operatorname{grad}\mathcal{F}^*(\boldsymbol{s}) = E_{P_{\boldsymbol{s}}}[\boldsymbol{Y}]$ for $\boldsymbol{s} \in \operatorname{int} D_{\mathcal{F}^*}$. Hence for each $\boldsymbol{s} \in \operatorname{int} D_{\mathcal{F}^*}$, the Kullback information of $P_{\boldsymbol{s}}$ with respect to μ is

$$\mathcal{K}(P_{\boldsymbol{s}}, \mu) = \boldsymbol{s}^t E_{P_{\boldsymbol{s}}}[\boldsymbol{Y}] - \mathcal{F}^*(\boldsymbol{s}) = \boldsymbol{s}^t \operatorname{grad}\mathcal{F}^*(\boldsymbol{s}) - \mathcal{F}^*(\boldsymbol{s}). \tag{16}$$

We will also use the *convex conjugate* $\mathcal{F}$ of $\mathcal{F}^*$

$$\mathcal{F}(\boldsymbol{y}) \triangleq \sup_{\boldsymbol{s}}\{\boldsymbol{s}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{s})\} \tag{17}$$

also called the *Cramér transform* of μ [30]. It is an l.s.c. convex function with domain $D_{\mathcal{F}}$. As $\mathcal{F}$ is itself l.s.c., the convex conjugate of $\mathcal{F}$ is $\mathcal{F}^*$ [37], which motivates the notation $\mathcal{F}^*$.

The strict convexity of $\mathcal{F}^*$ implies that $\mathcal{F}$ is *essentially smooth* (see, for instance, [28, p. 224]), which means that $\operatorname{int} D_{\mathcal{F}} \neq \emptyset$, $\mathcal{F}$ is differentiable over int $D_{\mathcal{F}}$, and $\mathcal{F}$ is *steep*, i.e., for any sequence $\{\boldsymbol{y}_n\}$ in $\operatorname{int} D_{\mathcal{F}}$ converging to a boundary point of $D_{\mathcal{F}}$

$$\lim_{n\mapsto\infty} \|\operatorname{grad}\mathcal{F}(\boldsymbol{y}_n)\| = +\infty.$$

For instance, a convex l.s.c. function whose domain is open and which is differentiable over its domain is steep (see [36, p. 87]).

3) Results:

Proposition 1: The Cramér transform $\mathcal{F}$ satisfies

$$\forall \boldsymbol{y} \in D_{\mathcal{F}}, \qquad \mathcal{F}(\boldsymbol{y}) \leq \min_{P\colon E_P[\boldsymbol{Y}]=\boldsymbol{y}} \mathcal{K}(P, \mu) \tag{18}$$

with equality if

$$\exists\, P_{\boldsymbol{s}} \in \xi_\mu \text{ such that } E_{P_{\boldsymbol{s}}}[\boldsymbol{Y}] = \boldsymbol{y}. \tag{19}$$

Moreover, if μ is maximal with steep log-Laplace transform, then a sufficient condition for (19) is that $\boldsymbol{y}$ belongs to the interior of the closed and convex hull of the support of μ, that is,

$$\boldsymbol{y} \in \operatorname{int}\{\operatorname{cc} S_\mu\}. \tag{20}$$

Sketch of the Proof: Checking the first inequality is straightforward: $\forall \boldsymbol{y} \in D_{\mathcal{F}}$ and $\forall \boldsymbol{s} \in D_{\mathcal{F}^*}$ let P be a probability measure of mean $\boldsymbol{y}$ and finite μ-entropy. P is absolutely continuous with respect to μ and to $P_{\boldsymbol{s}}$, because μ and $P_{\boldsymbol{s}}$ are equivalent. Using the Radon–Nikodim derivative of P with respect to $P_{\boldsymbol{s}}$ and (16), we derive

$$\begin{aligned}\mathcal{K}(P, \mu) &= \int \log\frac{dP}{dP_{\boldsymbol{s}}}\,dP + \int \log\frac{dP_{\boldsymbol{s}}}{d\mu}\,dP \\ &= \mathcal{K}(P, P_{\boldsymbol{s}}) + \int \bigl(\boldsymbol{s}^t\boldsymbol{u} - \mathcal{F}^*(\boldsymbol{s})\bigr)\,dP(\boldsymbol{u}) \\ &= \mathcal{K}(P, P_{\boldsymbol{s}}) + \boldsymbol{s}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{s}) \geq \boldsymbol{s}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{s}) \end{aligned} \tag{21}$$

and, as the last inequality is true for all $\boldsymbol{s}$

$$\mathcal{K}(P, \mu) \geq \sup_{\boldsymbol{s}}\{\boldsymbol{s}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{s})\} = \mathcal{F}(\boldsymbol{y})$$

which implies (18). The case of equality (19) is straightforward. Condition (20) derives from the following identity between sets:

$$\operatorname{int} D_{\mathcal{F}} = \operatorname{int}\{\operatorname{cc} S_\mu\}. \tag{22}$$

Under the hypotheses of Proposition 1, $(\operatorname{int} D_{\mathcal{F}}, \mathcal{F})$ and $(\operatorname{int} D_{\mathcal{F}^*}, \mathcal{F}^*)$ are Legendre transforms of each other [37] and

$$\operatorname{int}\{\operatorname{range}\{\operatorname{grad}\mathcal{F}^*\}\} = \operatorname{int} D_{\mathcal{F}}. \tag{23}$$

A proof of (22) and (23) can be found in [36, Sec. 9.1]. □

Proposition 2: If μ is a maximal probability measure, whose log-Laplace transform $\mathcal{F}^*$ is steep and satisfies

$$\mathbf{0} \in \operatorname{int} D_{\mathcal{F}^*} \tag{24}$$

and if the data satisfy the qualification constraint

$$\exists \boldsymbol{y} \in D_{\mathcal{F}} : \boldsymbol{H}\boldsymbol{y} = \boldsymbol{z}$$

then there is a unique solution to the problem

$$\min_{\boldsymbol{y}:\boldsymbol{H}\boldsymbol{y}=\boldsymbol{z}} \mathcal{F}(\boldsymbol{y}). \tag{25}$$

Proof: Proposition 2 can be derived from results on convex functions, see [37, Sec. 27] or from general results on Kullback–Leibler information [23], [38]. However, we give a simple and self-contained proof in the sequel. The assumptions on μ and $\mathcal{F}^*$ imply that $\mathcal{F}$ is strictly convex. To ensure that the minimum of $\mathcal{F}$ is attained we use the fact that the level-sets of $\mathcal{F}$ are compact and the Weierstrass theorem. We already know that these sets are closed because $\mathcal{F}$ is l.s.c. The fact that they are bounded is a consequence of assumption (24). If (24) holds, one can find a $\delta > 0$, such that the ball $\{\boldsymbol{s}\colon \|\boldsymbol{s}\| \leq \delta\}$ is a subset of $\operatorname{int} D_{\mathcal{F}^*}$. Then $c = \sup\{\mathcal{F}^*(\boldsymbol{s}),\ \|\boldsymbol{s}\| \leq \delta\} < \infty$, and, for every $\boldsymbol{y} \in \mathbb{R}^K$, using (17)

$$\begin{aligned}\mathcal{F}(\boldsymbol{y}) &\geq \sup\bigl\{\boldsymbol{s}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{s}),\ \|\boldsymbol{s}\| \leq \delta\bigr\} \\ &\geq \sup\bigl\{\boldsymbol{s}^t\boldsymbol{y},\ \|\boldsymbol{s}\| \leq \delta\bigr\} - c \geq \delta\|\boldsymbol{y}\| - c.\end{aligned}$$

For every $\alpha \in \mathbb{R}$, $\mathcal{F}(\boldsymbol{y}) \leq \alpha$ implies $\|\boldsymbol{y}\| \leq (\alpha + c)/\delta$ and the α-level set is bounded. □

Finally, we present sufficient conditions to obtain a duality result associated with both problems (9) and (25).

Proposition 3: If μ is a maximal probability measure, whose log-Laplace transform $\mathcal{F}^*$ is steep and satisfies $\mathbf{0} \in \operatorname{int} D_{\mathcal{F}^*}$, and if the data satisfy the *strong* qualification constraint

$$\exists \boldsymbol{y} \in \operatorname{int}\{\operatorname{cc} S_\mu\} : \boldsymbol{H}\boldsymbol{y} = \boldsymbol{z} \tag{26}$$

then the following primal-dual attainment equation is valid:

$$\begin{cases} \min\limits_{\boldsymbol{y}:\boldsymbol{H}\boldsymbol{y}=\boldsymbol{z}} \mathcal{F}(\boldsymbol{y}) = \mathcal{F}(\hat{\boldsymbol{y}}) = \mathcal{D}(\hat{\boldsymbol{\lambda}}) = \max\limits_{\boldsymbol{\lambda}} \mathcal{D}(\boldsymbol{\lambda}) \\ \text{with } \mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t \boldsymbol{z} - \mathcal{F}^*(\boldsymbol{H}^t\boldsymbol{\lambda}) \end{cases} \tag{27}$$

and where the primal solution $\hat{\boldsymbol{y}}$ satisfies

$$\hat{\boldsymbol{y}} = \operatorname{grad} \mathcal{F}^*(\hat{\boldsymbol{s}}) = E_{P_{\hat{\boldsymbol{s}}}}[\boldsymbol{Y}] \tag{28}$$

with $\hat{\boldsymbol{s}} = \boldsymbol{H}^t\hat{\boldsymbol{\lambda}}$, and $P_{\hat{\boldsymbol{s}}}$ gives the P^{ME} of problem (9). The dual criterion $\mathcal{D}(\boldsymbol{\lambda})$ is a concave function, which is strictly concave if $\operatorname{Ker} \boldsymbol{H}^t = \{\mathbf{0}\}$.

Proof: Using (22), assumption (26) is equivalent to $\exists \boldsymbol{y} \in \operatorname{int} D_{\mathcal{F}}$: $\boldsymbol{H}\boldsymbol{y} = \boldsymbol{z}$. It implies that the minimum in (27) cannot be attained at a boundary point of $D_{\mathcal{F}}$, by property of essential smoothness of $\mathcal{F}$, see [37, p. 252].

The rest of the proposition is an application of the Fenchel duality theorem [37]. Since $\hat{\boldsymbol{y}} \in \operatorname{int} D_{\mathcal{F}}$, a dual parameter $\hat{\boldsymbol{s}}$ can be associated to it by (28), and

$$\exists \hat{\boldsymbol{\lambda}} : \hat{\boldsymbol{s}} = \operatorname{grad} \mathcal{F}(\hat{\boldsymbol{y}}) = \boldsymbol{H}^t\hat{\boldsymbol{\lambda}}$$

by application of the standard result on Lagrange multipliers.

By (17) with $\boldsymbol{s} = \boldsymbol{H}^t\boldsymbol{\lambda}$, we have

$$\mathcal{F}(\hat{\boldsymbol{y}}) \geq \boldsymbol{\lambda}^t\boldsymbol{H}^t\hat{\boldsymbol{y}} - \mathcal{F}^*(\boldsymbol{H}^t\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{z} - \mathcal{F}^*(\boldsymbol{H}^t\boldsymbol{\lambda}).$$

The equality holds for $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$ and we define

$$\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{z} - \mathcal{F}^*(\boldsymbol{H}^t\boldsymbol{\lambda}).$$

If $\operatorname{Ker} \boldsymbol{H}^t = \{\mathbf{0}\}$, $\mathcal{D}$ is strictly concave and $\hat{\boldsymbol{\lambda}}$ is its unique minimizer. The rest follows by Proposition 1. □

D. Properties of the Reconstruction Process

We have presented a method to design a criterion for signal reconstruction purposes. We recall here the assumptions, discuss their meaning, and derive the main properties of the resulting reconstruction process, keeping in mind our original convex inverse problem (1), (2).

1) General Assumptions: Given the linear noisy problem

$$[\boldsymbol{A}, \boldsymbol{I}]\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{n}\end{bmatrix} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{z} \tag{29}$$

and the convex constraints

$$\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^K \quad \text{and} \quad \boldsymbol{n} \in \mathcal{B} \subset \mathbb{R}^N \tag{30}$$

the estimation method merely consists of estimating $(\boldsymbol{x}, \boldsymbol{n})$ by

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{n}}) = \operatorname{argmin} \mathcal{F}(\boldsymbol{x}, \boldsymbol{n}) \text{ subject to (29)}$$

where $\mathcal{F}$ is the Cramér transform of a reference probability measure μ defined on $\mathbb{R}^{K+N}$. We have also indicated that the corresponding dual problem is

$$\hat{\boldsymbol{\lambda}} = \operatorname*{argmax}_{\boldsymbol{\lambda}} \mathcal{D}(\boldsymbol{\lambda}) = \operatorname*{argmax}_{\boldsymbol{\lambda}} \left\{\boldsymbol{\lambda}^t\boldsymbol{z} - \mathcal{F}^*(\boldsymbol{H}^t\boldsymbol{\lambda})\right\}.$$

Let us first recall the conditions introduced above to ensure the existence, unicity, and dual form of the solution.

C1: μ is a maximal probability on $\mathbb{R}^{K+N}$;

C2: $\mathbf{0} \in \operatorname{int} D_{\mathcal{F}^*}$ and $\mathcal{F}^*$ is steep;

C3: $\exists \boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{n}) \in \operatorname{int}\{\mathcal{X} \times \mathcal{B}\} : \boldsymbol{H}\boldsymbol{y} = \boldsymbol{z}$.

Note that the condition $\operatorname{Ker} \boldsymbol{H}^t = \{\mathbf{0}\}$, which ensures the strict convexity of $\mathcal{D}(\boldsymbol{\lambda})$, is automatically satisfied, due to the special form of $\boldsymbol{H}$. The role of the reference measure is essential: it is the only way to incorporate prior information in the estimation process, and to ensure its well-posedness. In the sequel we postulate the independence of the noise and the object, and we take a product measure

C4: product measure: $\mu = \mu_{\boldsymbol{x}} \otimes \mu_{\boldsymbol{n}}$.

Finally, as far as the convex constraints (30) are concerned, we choose the reference measures $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{n}}$ such that

C5: supports of the reference measures satisfy

$$\operatorname{cc} S_{\mu_{\boldsymbol{x}}} = \mathcal{X} \quad \text{and} \quad \operatorname{cc} S_{\mu_{\boldsymbol{n}}} = \mathcal{B}.$$

2) Properties of the Reconstruction Process:

Property 1: As a consequence of C4, the Cramér and log-Laplace transforms are sum-separable

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{n}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n})$$

and

$$\mathcal{F}^*(\boldsymbol{\lambda}_{\boldsymbol{x}}, \boldsymbol{\lambda}_{\boldsymbol{n}}) = \mathcal{F}^*_{\boldsymbol{x}}(\boldsymbol{\lambda}_{\boldsymbol{x}}) + \mathcal{F}^*_{\boldsymbol{n}}(\boldsymbol{\lambda}_{\boldsymbol{n}}). \tag{31}$$

Using the observation equation and (27), the primal and dual criteria are

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$$

and

$$\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{z} - \mathcal{F}^*_{\boldsymbol{x}}(\boldsymbol{A}^t\boldsymbol{\lambda}) - \mathcal{F}^*_{\boldsymbol{n}}(\boldsymbol{\lambda}) \tag{32}$$

and the primal-dual relation is

$$\hat{\boldsymbol{x}} = \operatorname{grad} \mathcal{F}^*_{\boldsymbol{x}}(\boldsymbol{A}^t\hat{\boldsymbol{\lambda}}). \tag{33}$$

Equation (31) is a direct consequence of definitions (15) and (17). Thus we have obtained the compound criterion of the general form (4), together with its dual formulation. It is possible to take advantage of the two dual formulations of the optimization problem. For instance, the dual criterion is particularly useful when the Cramér transforms of the measures do not have an analytic form. In such cases, the dual formulation can be used to solve the implicit primal problem. Therefore, the method only requires an explicit form of the log-Laplace transform $\mathcal{F}^*$.

Property 2—Domain Constraint: Under assumptions C1, C2, C4, and C5

$$\operatorname{int} D_{\mathcal{F}} = \operatorname{int} \mathcal{X} \times \operatorname{int} \mathcal{B}$$

and

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{n}) = +\infty \text{ if } \boldsymbol{x} \notin \mathcal{X} \text{ or } \boldsymbol{n} \notin \mathcal{B}.$$

This property has been documented in the proof of Proposition 1, see (22). The proof that $\mathcal{F}$ is infinite outside the domain

of the constraints can be found in [36]. The behavior of $\mathcal{F}$ on the boundary of $\mathcal{X}$ and $\mathcal{B}$ depends on the existence of masses on the boundary, see [30] and [36].

The behavior of the primal and dual criteria inside their domain is also very attractive:

Property 3: Under assumptions C1, C2, C4, and C5, $\mathcal{F}$ is continuously differentiable and strictly convex on $\operatorname{int} D_{\mathcal{F}}$. If $\boldsymbol{y}$ is a finite boundary point of $D_{\mathcal{F}}$, and $\{\boldsymbol{y}_n\}$ a sequence of points of $\operatorname{int} D_{\mathcal{F}}$ converging to $\boldsymbol{y}$

$$\lim_{n\to\infty} \mathcal{F}(\boldsymbol{y}_n) = +\infty, \qquad \text{if } \boldsymbol{y} \notin D_{\mathcal{F}}$$
$$\lim_{n\to\infty} \|\operatorname{grad} \mathcal{F}(\boldsymbol{y}_n)\| = +\infty.$$

The first property stems from the l.s.c., while the second is the steepness of $\mathcal{F}$. This limiting behavior of the cost function is suited for optimization by steepest gradient descent: during the optimization process, the current object is "pushed away" from the boundary of the domain of the constraint (note the importance of the strong qualification condition C3).

Property 4: $\mathcal{F}$ is a positive function. It takes its minimum value of zero at the unique point $\boldsymbol{y}=\boldsymbol{m}_{\boldsymbol{y}}$, i.e., $\boldsymbol{x}=\boldsymbol{m}_{\boldsymbol{x}}$, $\boldsymbol{n}=\boldsymbol{m}_{\boldsymbol{n}}$, the expectations of the reference measures $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{n}}$.

As a consequence of this property, we can consider the primal criterion as a discrepancy measure with respect to the reference mean value, and note it $\mathcal{F}(\boldsymbol{y}|\boldsymbol{m}_{\boldsymbol{y}})$ or, using the separability and (29)

$$\mathcal{F}(\boldsymbol{y}|\boldsymbol{m}_{\boldsymbol{y}}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}_{\boldsymbol{x}}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z}-\boldsymbol{A}\boldsymbol{x}|\boldsymbol{m}_{\boldsymbol{n}}). \tag{34}$$

Hence the criterion appears as a tradeoff between two discrepancy measures penalizing the difference between, respectively, the object and the prior mean $\boldsymbol{m}_{\boldsymbol{x}}$; the residuals and the noise mean $\boldsymbol{m}_{\boldsymbol{n}}$.

We have already seen that for every member of the exponential family ξ_μ, having an expected value $\boldsymbol{y} \in \operatorname{int} D_{\mathcal{F}}$, say $P_{\boldsymbol{s}}$ with $\boldsymbol{s} = \operatorname{grad} \mathcal{F}(\boldsymbol{y})$, the μ-entropy may be written as

$$\mathcal{K}(P_{\boldsymbol{s}}, \mu) = \boldsymbol{s}^t\boldsymbol{y} - \mathcal{F}^*(\boldsymbol{s}) = \mathcal{F}(\boldsymbol{y}|\boldsymbol{m}_{\boldsymbol{y}}).$$

It is possible to give to this relation an interesting extension by considering the measure of the discrepancy between two members of $\operatorname{int} D_{\mathcal{F}}$. For this we define

$$\mathcal{F}(\boldsymbol{y}_1|\boldsymbol{y}_2) \triangleq \mathcal{K}(P_{\boldsymbol{s}_1}, P_{\boldsymbol{s}_2}), \quad \boldsymbol{s}_i = \operatorname{grad} \mathcal{F}(\boldsymbol{y}_i), \qquad i = 1, 2.$$

A calculation similar to (21) leads to

$$\begin{aligned}\mathcal{K}(P_{\boldsymbol{s}_1}, P_{\boldsymbol{s}_2}) &= \mathcal{K}(P_{\boldsymbol{s}_1}, \mu) - \boldsymbol{s}_2^t\boldsymbol{y}_1 + \mathcal{F}^*(\boldsymbol{s}_2) \\ &= \mathcal{F}(\boldsymbol{y}_1|\boldsymbol{m}_{\boldsymbol{y}}) - \mathcal{F}(\boldsymbol{y}_2|\boldsymbol{m}_{\boldsymbol{y}}) - \boldsymbol{s}_2^t(\boldsymbol{y}_1 - \boldsymbol{y}_2).\end{aligned}$$

Thus

$$\mathcal{F}(\boldsymbol{y}_1|\boldsymbol{y}_2) = \mathcal{F}(\boldsymbol{y}_1|\boldsymbol{m}_{\boldsymbol{y}}) - \mathcal{F}(\boldsymbol{y}_2|\boldsymbol{m}_{\boldsymbol{y}}) - (\operatorname{grad} \mathcal{F}(\boldsymbol{y}_2))^t(\boldsymbol{y}_1 - \boldsymbol{y}_2). \tag{35}$$

Equation (35) establishes the link between the entropic cost functions presented here and Bregman divergences [39]. Moreover, it will be useful when discussing a possibility of changing the initial guess $\boldsymbol{m}_{\boldsymbol{y}}$ in the estimation procedure, see Section IV-B1). We use it here to demonstrate the following property.

Property 5: Under assumptions C1–C5, let $\hat{\boldsymbol{y}}$ denote the reconstruction using criterion $\mathcal{F}$. For every $\boldsymbol{y} \in \operatorname{int} D_{\mathcal{F}}$ such that $\boldsymbol{H}\boldsymbol{y} = \boldsymbol{z}$, the following "Pythagorean" equation holds:

$$\mathcal{F}(\boldsymbol{y}|\boldsymbol{m}_{\boldsymbol{y}}) = \mathcal{F}(\boldsymbol{y}|\hat{\boldsymbol{y}}) + \mathcal{F}(\hat{\boldsymbol{y}}|\boldsymbol{m}_{\boldsymbol{y}}). \tag{36}$$

Indeed, choosing $\boldsymbol{y}_1 = \boldsymbol{y}$ and $\boldsymbol{y}_2 = \hat{\boldsymbol{y}}$ in (35), and using the fact that the dual optimal parameter is $\hat{\boldsymbol{s}} = \boldsymbol{H}^t\hat{\boldsymbol{\lambda}}$, we have

$$\begin{aligned}\mathcal{F}(\boldsymbol{y}|\hat{\boldsymbol{y}}) &= \mathcal{F}(\boldsymbol{y}|\boldsymbol{m}_{\boldsymbol{y}}) - \mathcal{F}(\hat{\boldsymbol{y}}|\boldsymbol{m}_{\boldsymbol{y}}) - \hat{\boldsymbol{\lambda}}^t\boldsymbol{H}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \\ &= \mathcal{F}(\boldsymbol{y}|\boldsymbol{m}_{\boldsymbol{y}}) - \mathcal{F}(\hat{\boldsymbol{y}}|\boldsymbol{m}_{\boldsymbol{y}}).\end{aligned}$$

The last property is called "directed orthogonality" in [2] (see also [5]). Actually, it is a property of the minimization of Kullback information under linear constraints [22], [23], which is also associated with all Bregman divergences. Further interesting geometrical consequences of (36) can be found in [3] and [6].

Finally, the following special case will be useful for many applications:

Property 6: If $\mathcal{X}$ is a product set such as (3) so that

$$\mathcal{X} = \{\boldsymbol{x} : \quad a_j \le x_j \le b_j, \quad 1 \le j \le K\}$$

and $\mu_{\boldsymbol{x}}$ a product measure

$$\mu_{\boldsymbol{x}} = \bigotimes_{j=1}^{K} \mu_{x_j}$$

with each μ_{x_j} having a support whose closed and convex hull is $[a_j, b_j]$, then $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{x}}^*$ are sum-separable

$$\mathcal{F}_{\boldsymbol{x}} = \sum_{j=1}^{K} f_j \qquad \mathcal{F}_{\boldsymbol{x}}^* = \sum_{j=1}^{K} f_j^*$$

and the closure of $D_{\mathcal{F}}$ is $\mathcal{X}$. The primal-dual relationship is a (generally nonlinear) component-wise transformation in $\mathbb{R}^K$

$$\hat{x}_j = (f_j^*)'(\hat{s}_j), \qquad 1 \le j \le K, \text{ with } \hat{\boldsymbol{s}} = \boldsymbol{A}^t\hat{\boldsymbol{\lambda}}_{\boldsymbol{x}}.$$

III. Examples of Criteria

In this section, we give, in the form of a table with comments and references, several examples of reference measures which encode different type of prior informations and convex constraints. For each measure, we give the log-Laplace transform $\mathcal{F}^*(\boldsymbol{s})$, that defines the dual criterion, and, when explicit, the primal criterion $\mathcal{F}(\boldsymbol{x})$. The primal-dual relations are given by the derivative of the log-Laplace transforms. Except in the case of Gaussian measures, we consider separable reference measures yielding separable criteria. These criteria are given here in terms of a general object $\boldsymbol{x}$ and a measure $\mu_{\boldsymbol{x}}$, not necessarily related to the "object" $\boldsymbol{x}$ in the inverse problem formulation.

Next, we illustrate the construction of criteria within the entropic frame while studying a particular family of reference measure, that give rise to interesting special cases.

A. Table of Criteria

Table I (at the top of the following page) presents several reference measures and criteria.

TABLE I
EXAMPLES OF REFERENCE MEASURES μ, ASSOCIATED LOG-LAPLACE TRANSFORMS $\mathcal{F}^*$ (FROM WHICH DUAL CRITERIA ARE DERIVED) AND CRAMÉR TRANSFORMS $\mathcal{F}$ (PRIMAL CRITERIA). THE SETS OF CONSTRAINTS ARE INDICATED IN FIRST COLUMN (THE NOTATION $\mathcal{X}_j$ REFERS TO THE CASE OF SEPARABLE CONVEX SETS $\mathcal{X} = \bigotimes_{j=1}^{K} \mathcal{X}_j$). BOTTOM PART OF THE TABLE SHOWS COMPONENTS OF SEPARABLE MEASURES AND CRITERIA (FOR INSTANCE, IN SECOND COLUMN, $\mathcal{F}^*(\boldsymbol{s}) = \sum_{j=1}^{K} f_j^*(\boldsymbol{s}_j)$)

$\mathcal{X}$	μ	$\mathcal{F}^*(\boldsymbol{s})$	$\mathcal{F}(\boldsymbol{x})$	*comments/references*
$\mathbb{R}^K$	$\mathcal{N}(\boldsymbol{m}, \boldsymbol{R}_x)$	$\frac{1}{4}\boldsymbol{s}^t\boldsymbol{R}_x\boldsymbol{s} + \boldsymbol{s}^t\boldsymbol{m}$	$(\boldsymbol{x}-\boldsymbol{m})^t\boldsymbol{R}_x^{-1}(\boldsymbol{x}-\boldsymbol{m})$	see references [30, 7, 9]
$\mathbb{R}^K$	$\gamma\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_1) + (1-\gamma)\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_2)$	$\log\left(\gamma \exp\left(\frac{\boldsymbol{s}^t\boldsymbol{R}_1\boldsymbol{s}}{2}\right) + (1-\gamma)\exp\left(\frac{\boldsymbol{s}^t\boldsymbol{R}_2\boldsymbol{s}}{2}\right)\right)$	not explicit	applied to sparse spike trains deconvolution [16]
$\mathcal{X}_j$	μ_j	$f_j^*(s_j)$	$f_j(x_j)$	*comments/references*
$\mathbb{R}^+$	Poisson(m_j)	$m_j(\exp(s_j) - 1)$	$x_j \log\frac{x_j}{m_j} - x_j + m_j$	see following subsection
$\mathbb{R}^+$	$\Gamma(\alpha_j, \beta_j)$ (mean : $m_j = \alpha_j/\beta_j$)	$\beta_j \log\left(\frac{\alpha_j}{\alpha_j - s_j}\right)$	$\beta_j\left[\left(\frac{x_j}{m_j} - 1\right) - \log\left(\frac{x_j}{m_j}\right)\right]$	Itakura-Saito measure, used in spectrum analysis [40]; case $m_j = 1$ is the Burg entropy [41, 42, 7]
$\mathbb{R}^+$	$\gamma\Gamma(\alpha_1, \beta_1) + (1-\gamma)\Gamma(\alpha_2, \beta_2)$	$\log\left(\gamma\frac{\beta_1^{\alpha_1}}{(\beta_1 - s_j)^{\alpha_1}} + (1-\gamma)\frac{\beta_2^{\alpha_2}}{(\beta_2 - s_j)^{\alpha_2}}\right)$	not explicit	used in spiky positive signals reconstruction [16], *e.g.* in high resolution spectral analysis
$\mathbb{R}^+$	law of $X = \sum_{l=1}^{L} Y_{j,l}$ with L a Poisson(λ_j) random variable and $Y_{j,l} \sim \Gamma(\alpha_j, \beta_j)$	$\lambda_j\left(\frac{\beta_j}{\beta_j - m_j s_j}\right)^{\beta_j} - \lambda_j$	$\lambda_j\left\{\beta_j\frac{x_j}{m_j} + 1 - (\beta_j+1)\left(\frac{x_j}{m_j}\right)^{\frac{\beta_j}{\beta_j+1}}\right\}$	see following subsection
$[a_j, b_j]$	uniform measure	$\log\left(\frac{e^{a_j s_j} - e^{b_j s_j}}{s_j}\right)$	not explicit	used in crystallographic applications [7]; see also [9]
$[a_j, b_j]$	Bernoulli measure: $\alpha_j\delta(x_j - a_j) + (1-\alpha_j)\delta(x_j - b_j)$	$\log(\alpha_j \exp(s_j a_j) + (1-\alpha_j)\exp(s_j b_j))$	$\frac{x_j - a_j}{b_j - a_j}\log\left(\frac{x_j - a_j}{1-\alpha_j}\right) + \frac{b_j - x_j}{b_j - a_j}\log\left(\frac{b_j - x_j}{\alpha_j}\right) - \log(b_j - a_j)$	used in crystallographic applications; case $a_j = 0$ and $b_j = 1$ is the Fermi-Dirac entropy, see [7, 9]

B. Poissonized Sum Distributions

An interesting family of reference measures, proposed in [12], is obtained with a Poissonized sum of random variables. We consider a random variable X_j, defined as the sum of L_j independent random variables $Y_{j,l}$, $l = 1 \cdots L_j$, with common distribution Q; L_j being a Poisson variable with parameter λ_j

$$X_j = \sum_{l=1}^{L_j} Y_{j,l}.$$

This model may be seen as an image formation model. The image is a plane divided into K cells, and its intensity results from the fall of a random number of quanta, following a Poisson distribution, of mean parameter λ_j at site j. The intensity of each quanta is variable, and governed by the distribution Q.

Suppose now that the closed and convex hull of the support of Q is $\mathbb{R}^+$. Let μ_j denote the law of X_j and define μ as the product

$$\mu = \bigotimes_{j=1}^{K} \mu_j.$$

A simple calculation shows that the log-Laplace transform $\mathcal{F}^*$ of μ depends on the Laplace transform $\mathcal{L}_Q$ of Q through

$$\mathcal{F}^*(\boldsymbol{s}) = \sum_{j=1}^{K} \lambda_j \mathcal{L}_Q(s_j) - \lambda_j.$$

Poisson Distribution: We first apply the construction in the special case of a deterministic $Y_{i,l}$, and we take simply $Q(Y_{i,l}) = \delta(Y_{i,l} - 1)$. The probability distribution of any sample $\boldsymbol{x} = [x_1, \cdots, x_K]$ is thus a Poisson distribution $\mu_{\boldsymbol{x}}$ with mean parameter λ_j. We take here the notation $\boldsymbol{\lambda} = \boldsymbol{m}$. Then the primal criterion $\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m})$ is obtained as the convex conjugate of $\mathcal{F}_{\boldsymbol{x}}^*$, that is, the Cramér transform of $\mu_{\boldsymbol{x}}$

$$\begin{cases} \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{s}) = \sum_{j=1}^{K} m_j(\exp(s_j) - 1) \\ \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}) = \sum_{j=1}^{K}\left\{x_j \log\frac{x_j}{m_j} - x_j + m_j\right\}. \end{cases}$$

The functional $\mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m})$ is none other than the generalized cross entropy (see [5] and [43]) between $\boldsymbol{x}$ and $\boldsymbol{m}$. This is the "$x \log x$" entropy (in reference to the title of [44]). The vector $\boldsymbol{m}$ is an initial guess for the sought object and is often chosen as a flat object, or as the result of some previous experiment. The "$x \log x$" entropic regularization is often termed and considered as an information theory based reconstruction method and has been motivated by several axiomatic works [3]–[5], [43].

This kind of regularization has been used with some success in several applied problems, such as astronomy, tomography, RMN, and spectrometry. As far as the positivity constraint is concerned, criteria like Section III-B built upon logarithmic expressions, ensure the positivity [45]. Another well-known example is the "$\log x$" entropy used in spectral analysis, see Table I.

Poissonnized Sum of Gamma Distributions: Another interesting application is the case of the Poissonnized sums of Gamma distributions [12]

$$d\mu_i(x_i) = \frac{\beta_i^{\beta_i}}{m_i^{\beta_i}\Gamma(\beta_i)} e^{-(\beta_i/m_i)x_i} x_i^{\beta_i - 1}\, dx_i.$$

The log-Laplace transform is

$$\mathcal{F}^*_{\boldsymbol{x}}(\boldsymbol{s}) = \sum_{j=1}^{K} \lambda_j \left(\frac{\beta_j}{\beta_j - m_j s_j} \right)^{\beta_j} - \lambda_j$$

and

$$\mathcal{F}(\boldsymbol{x}|\boldsymbol{m}) = \sum_{j=1}^{K} \lambda_j \left\{ \beta_j \frac{x_j}{m_j} + 1 - (\beta_j + 1) \left(\frac{x_j}{m_j} \right)^{\beta_j/(\beta_j+1)} \right\}. \tag{37}$$

This expression can be related to the entropy criteria introduced by Jones and Byrne, (see [3, Example 4], and [2]). In order to highlight the link with other studies we let $\beta_j = \beta, \forall\, j$ and introduce the new parameter $\gamma = \beta/(\beta+1)$. The criterion then becomes

$$\mathcal{F}(\boldsymbol{x}|\boldsymbol{m}) = \sum_{j=1}^{K} \frac{f_j}{(\gamma+1)m_j^{\gamma}} \left(-x_j^{\gamma} + \gamma m_j^{\gamma-1} x_j \right)$$

with $0 < \gamma < 1$. The latter expression is easily related (up to a constant) to Csiszár's "projection rule" in [5, eq. (3.7) (third line)]. Such criteria have been used in certain applied fields. For instance, in radioastronomy, Narayan and Nityananda have demonstrated in [45] the efficiency of the criterion (37) with $\beta = 1$, that is, the square-root criterion

$$\mathcal{F}(\boldsymbol{x}|\boldsymbol{m}) = \sum_{j=1}^{K} f_j \left(\frac{x_j}{m_j} + 1 - 2\sqrt{\frac{x_j}{m_j}} \right) \tag{38}$$

compared to the usual entropy criteria "$\log x$" and "$x \log x$." It is interesting to quote these authors [45, p. 144] who called $I^{1/2}$ the criterion (38):

"The success of $I^{1/2}$, which has no information-theoretic backing (no logarithms!), is a strong point in favor of the penalty function interpretation mentioned in Section II-V." We think that it is now clear to the reader that the square root criterion (38) does possess an "information-theoretic backing," as others which do not exhibit logarithms.

IV. Solving Real Problems

So far, we have described the principles of our use of entropy and several examples of reference measures which lead to interesting criteria. We shall now be concerned with the practical use of these criteria for real data inversion problems.

A. Criteria for the Noisy Problem

The resolution of an inverse problem begins by the specification, or the choice, of the reference measures. The additive noise has been taken into account through the use of an extended vector $\boldsymbol{y}^t = [\boldsymbol{x}^t, \boldsymbol{n}^t]$ and a separable reference measure. Concerning the noise term, we have found it useful to encode in the reference measure $\mu_{\boldsymbol{n}}$ the known characteristics of the noise. For instance, in the case of noise with known (or previously estimated), first and second-order characteristics, we take $\mathcal{B} = \mathbb{R}^K$ and a Gaussian distribution for $\mu_{\boldsymbol{n}}$, and

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}_{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{m}_{\boldsymbol{n}}^t)\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{m}_{\boldsymbol{n}})$$

and we find a "logical" term (the log-likelihood function) arising from the Gaussian behavior of the noise. In the case of non-Gaussian noise, the resulting term will be the Cramér transform of the noise reference measure. In the case of an additive Poisson perturbation of known intensity λ_j, we obtain

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}_{\boldsymbol{x}}) + \sum_{j=1}^{K} [\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}]_j \log \frac{[\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}]_j}{\lambda_j} - [\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}]_j + \lambda_j.$$

Here again the second term penalizes a "distance" between the residual and the noise mean. Observe that in the practical optimization, it is not necessary to ensure that $[\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}]_j \geq 0$, because of the steepness of $\mathcal{F}_{\boldsymbol{n}}$, see Property 3 in Section II-D2.

In the case of Poisson observations, such as in optical imagery, the data are Poisson-distributed with intensity $\boldsymbol{A}\boldsymbol{x}$, and therefore the "noise" is better described as a multiplicative process. It is possible to account for this situation in the form of an additive noise distributed as *centered* Poisson variables with a variable parameter. The general form of the goodness-of-fit term is

$$\mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}) = \sum_{j=1}^{K} [\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\lambda}]_j \log \frac{[\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\lambda}]_j}{\lambda_j} - [\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}]_j. \tag{39}$$

Using $\boldsymbol{\lambda} = \boldsymbol{z}$ is a good approximation for high photon count. It is also interesting to consider the choice $\boldsymbol{\lambda} = \boldsymbol{A}\boldsymbol{x}$, although we depart here from the assumption of independence between the object and the noise. However, formally replacing $\boldsymbol{\lambda}$ by $\boldsymbol{A}\boldsymbol{x}$ in (39) gives (up to a constant)

$$\sum_{j=1}^{K} -z_j \log([\boldsymbol{A}\boldsymbol{x}]_j) + [\boldsymbol{A}\boldsymbol{x}]_j$$

which is exactly the Poisson log-likelihood.

Other approaches can be used in order to account for the statistic of the observation noise. The entropic approach described here can always be used to extend the regularization term $\mathcal{F}_{\boldsymbol{x}}$, and, for instance, a "nonentropic" function can be added to it, or $\mathcal{F}_{\boldsymbol{x}}$ can be minimized subject to a set constraint defined using the noise distribution.

B. Choice of the Object Reference Measure

The choice of the reference measure is an important and difficult issue, which is reminiscent of selection of the prior distribution in the Bayesian setting, and is also an open problem. In our opinion, there is no general and definitive answer to this question. Three essential strategies and arguments, not exclusive, may be applied depending on the particular problem at hand. First, when the solution must belong to a convex set $\mathcal{C}_{\boldsymbol{y}}$, the reference measure should satisfy condition C5. Second, the measure can be derived using a "physical" model of the object formation. Third, a measure can be built, as a reasonable representation of the processes under study (for instance, the mixture distributions for spiky signals).

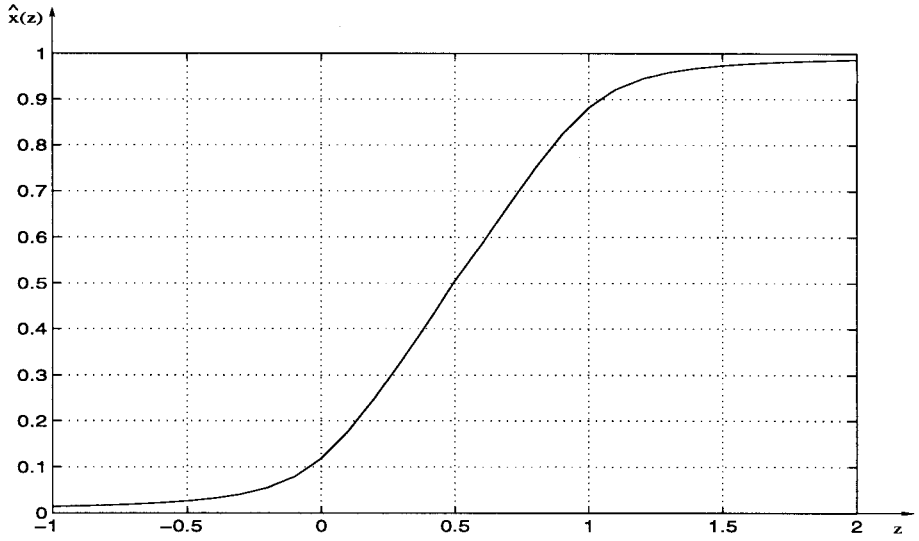


Fig. 1. "Transfer function" in the bounded case, that is, $\hat{x}(z) = \operatorname{argmin}_x \mathcal{F}_x(x) + \alpha\|z - x\|^2$, where $\mathcal{F}_x$ is built using a uniform reference over $[0, 1]$, forcing $\hat{x}(z)$ to belong to $[0, 1]$.

When several measures are possible, one should derive the "best" reconstruction for the problem at hand from knowledge about: 1) the desired behavior of the sought object and 2) the possibilities offered by the general framework of reconstruction which is used. We intend to give here information about 2) by illustrating some effects of changes of reference measure.

1) Changes of the Prior Guess: The criteria $\mathcal{F}(\boldsymbol{x}|\boldsymbol{m})$ have been interpreted as discrepancy measures between $\boldsymbol{x}$ and the prior guess $\boldsymbol{m}$, the mean of the reference measure. It is useful to be able to modify $\boldsymbol{m}$ in order to adapt, or refine, the procedure to another problem, while keeping the same form of criterion. The Bregman distances (35) are designed for this task. Suppose that a criterion $\mathcal{F}(\boldsymbol{x}|\boldsymbol{m})$ is associated with the reference measure μ and default object $\boldsymbol{m}$. In order to change the default object $\boldsymbol{m}$ into a new object $\boldsymbol{m}_1$, one can choose as the new reference measure the distribution P_1 in the exponential family ξ_μ generated by μ, with expectation parameter $\boldsymbol{m}_1$ and natural parameter $\boldsymbol{t}$. Then, the new primal criterion is derived using (35)

$$\mathcal{F}(\boldsymbol{x}|\boldsymbol{m}_1) = \mathcal{F}(\boldsymbol{x}|\boldsymbol{m}) - \mathcal{F}(\boldsymbol{m}_1|\boldsymbol{m}) - (\operatorname{grad}\mathcal{F}(\boldsymbol{m}_1))^t(\boldsymbol{m}_1 - \boldsymbol{x}).$$

Concerning the dual criterion, one can readily verify that the log-Laplace transform $\mathcal{F}^*_{P_1}$ of P_1 is simply related to the log-Laplace transform $\mathcal{F}^*_\mu$ of μ by

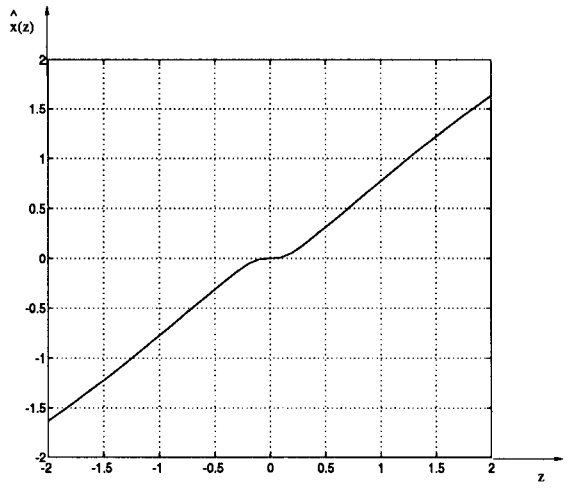
$$\mathcal{F}^*_{P_1}(\boldsymbol{s}) = \mathcal{F}^*_\mu(\boldsymbol{s} + \boldsymbol{t}) - \mathcal{F}^*_\mu(\boldsymbol{t}).$$

Hence changing the initial guess reduces to the addition of a linear term in the primal criterion, or equivalently, to shift the dual criterion.

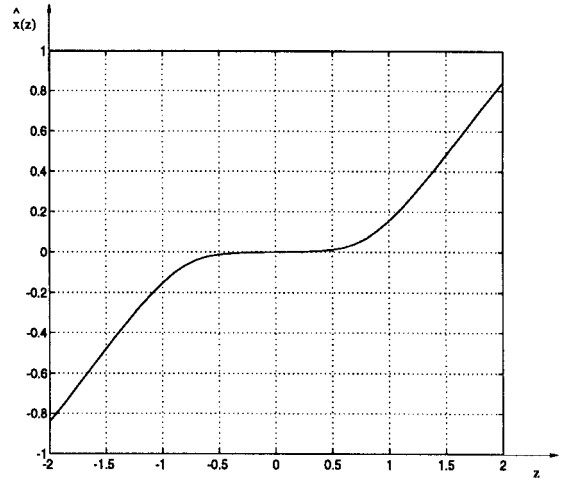
In order to reduce the bias toward the default object $\boldsymbol{m}$, some authors have proposed an iterative process, where the result of one inversion step is used as the new prior guess in the next inversion. Though the theoretical properties of this procedure have not, to our knowledge, been investigated, improved results have been reported, in [46] for instance.

2) Tuning the Nonlinear Effects: Nonquadratic regularization is characterized by nonlinear behavior of the reconstruction process, that allows, for instance, a convex constraint to be respected, and more generally, to go beyond the limitations of linear reconstruction. There have been several studies on "superresolution properties" of some nonlinear inversion methods, especially those deriving from entropic regularization. References can be found in physics (among others [7] and [45]) and in applied mathematics (see, for example, [11] and [47]). In [11], the superresolution phenomenon is linked to the nonlinear primal-dual relationship (33). Let us also mention the operational study of entropic reconstruction methods done by Narayan and Nityananda in [45], who have described the peak-sharpening and sidelobe-attenuating effects of these nonquadratic regularizers. Narayan and Nityananda argued that, from a pragmatic point of view, these effects were the real interest of entropic criteria if they were recognized and correctly used.

We can illustrate such nonlinear behavior using the notion of "*transfer function*" of the reconstruction process. We use the simple model $z = x + n$ (that is, the observation is scalar, and $\boldsymbol{A} = 1$), and define $\hat{x} = \operatorname{argmin}_x \mathcal{F}_{\boldsymbol{x}}(x|m) + \alpha(z - x)^2$. Then, we plot the reconstruction $\hat{x}$ as a function of the "observation" z. Quadratic regularization, $\mathcal{F}_{\boldsymbol{x}}(x|m) = x^2$, would give a linear estimate, which is simply $\hat{x} = \alpha/(1 - \alpha)z$. Thus $\hat{x}(z)$ has a constant slope. In contrast, reconstructions obtained with entropic criteria using non-Gaussian reference measures exhibit a variable slope: Fig. 1 shows the result with a bounded uniform measure, illustrating the nonlinear effect associated to the convex domain constraint. In the case of a measure defined as a mixture of Gaussian measures, two "*transfer functions*," in the sense precised above, are given in Fig. 2(a) and (b) for two different values of the mixture parameter. The nonlinear effect here is a *threshold* that separates the signal z into weak components, which are attenuated and large ones, which are respected. This effect is well adapted to the reconstruction of signals composed of rare spikes over a weak noise background,



(a)



(b)

Fig. 2. "Transfer functions" in the case of a mixture of Gaussian; $\hat{x}(z) = \text{argmin}_x \mathcal{F}_x(x) + \alpha\|z - x\|^2$. There is here a nonlinear effect, like a *threshold* that separates the signal z into weak and large components. (a) First set of parameters. (b) Second set of parameters. See [31] for an application of mixture of Gaussian to spike sparse train deconvolution.

as may occur in seismic data processing (see [16] and the reference therein). Varying the mixture parameter is equivalent to varying the threshold.

C. Some Extensions of the Entropic Setting

1) Nonlinear Operator: Thus far, we considered that the observations are a *linear* and noisy transform of the object, and we derived the form of the regularized criterion, together with its dual formulation. The case of a nonlinear operator is a straightforward extension, when one recognizes that the entropy functional is unchanged

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{n}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}_{\boldsymbol{x}}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{n}|\boldsymbol{m}_{\boldsymbol{n}}).$$

Then, incorporating the observation equation $\boldsymbol{z} = \boldsymbol{A}(\boldsymbol{x}) + \boldsymbol{n}$, one simply obtains

$$\mathcal{F}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}_{\boldsymbol{x}}) + \mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{A}(\boldsymbol{x})|\boldsymbol{m}_{\boldsymbol{n}}).$$

Despite the apparent simplicity of the last relation, the problem is far more complicated: the functional $\mathcal{F}_{\boldsymbol{n}}(\boldsymbol{z} - \boldsymbol{A}(\boldsymbol{x})|\boldsymbol{m}_{\boldsymbol{n}})$ is often no longer convex, and therefore the solution may not be unique and practical optimization may be much more difficult. An example of such approach can be found in [14] and [35], where the observations are modulus of some Fourier coefficients of the object.

2) Correlations: Correlations between the components of the objects are difficult to introduce directly in the entropic construction: the main difficulty being the computation of the log-Laplace transform of a nonseparable measure.

Indirect approaches can be found in [48] and [49]. Our approach [17] consists of the introduction of a "hidden" object $\boldsymbol{w}$ that is a linear transform of $\boldsymbol{x}$. We take, for example, $\boldsymbol{w} = \boldsymbol{D}\boldsymbol{x}$, where $\boldsymbol{D}$ is a differentiation matrix: $w_i = x_i - x_{i-1}$. We then use a reference measure for $\boldsymbol{w}$, encoding, for example, concentration, boundness, or a "spiky" character of the first-order derivatives. Then, using a separability assumption, one obtains the new entropy functional

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{w}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}_1) + \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{w}|\boldsymbol{m}_2).$$

Finally, using the constraint $\boldsymbol{w} = \boldsymbol{D}\boldsymbol{x}$, with a first-order differentiation matrix (for example), we have

$$\mathcal{F}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{m}) + \sum_i f_{w_i}(x_i - x_{i-1}|m_{2,i}).$$

This approach gives very interesting results [17]. Actually, the resulting criteria can be compared to convex potentials arising in a Markovian context. Other approaches can be found in [13] and [19].

D. Examples

1) Image Deblurring: An image FOC-f/96 (*faint object camera*) of Supernova SN1987A which exploded in February 1987 is given in Fig. 3. Data in Fig. 3(a) are blurred by the large impulse response (Fig. 3(b)) of the Hubble Space Telescope (HST) (before its correction in January 1994). The transfer matrix $\boldsymbol{A}$ was derived from this known impulse response. At this point, the problem is to deconvolve the data in Fig. 3(a) using the point-spread function (psf) (Fig. 3(b)). Because the psf has a very large support, spreading the energy in a large domain (although the main feature of the psf is narrow), this deconvolution problem is ill-conditioned, and a simple inversion, using for instance a generalized inverse, is unsatisfactory and does not provide more information than the direct data. In this problem, the goal is to be able to precisely measure distances between the features, and the repartition of energy in the reconstruction.

In order to cope with this problem, one is led to complete the data with some constraint or *a priori* information. A first kind of information is positivity, since the data are an energy distribution. In Fig. 3 (d), we applied a standard maximum-entropy "$x \log x$" reconstruction for positive objects, using the criterion

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}\|^2 + \gamma \sum_{i=1}^{K} \left(x_i \log \frac{x_i}{m_i} - x_i + m_i \right)$$

with $\boldsymbol{m}$ chosen as a "flat" object. For the reconstruction of astronomical objects, experimentators sometimes use a support

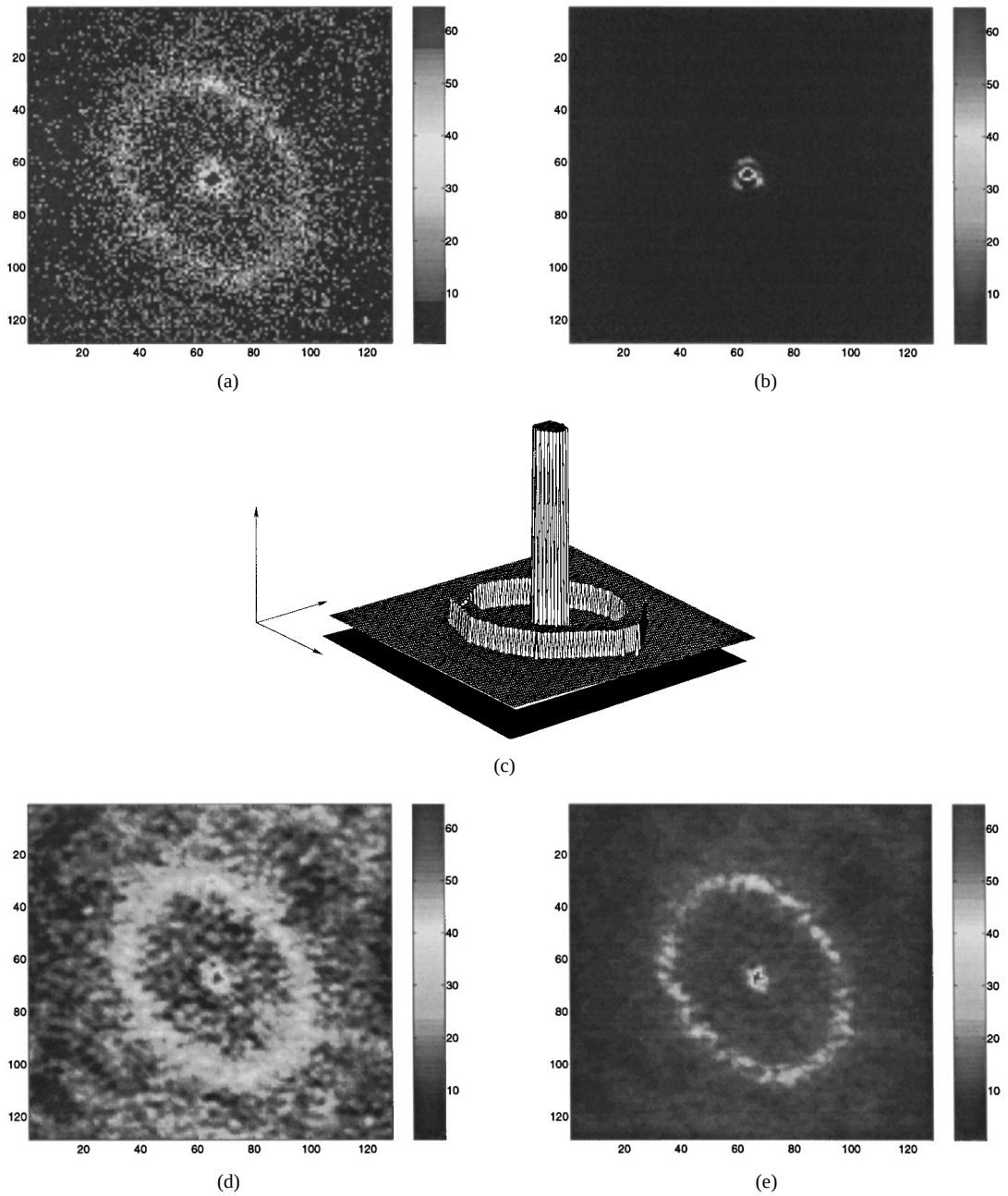


Fig. 3. Deconvolution of the data of the Supernova SN1987a, before its correction in January 1994: (a) data, (b) point-spread function, (c) higher bounds on the intensity of the object, deduced from the data, (d) "$x \log x$" reconstruction (positivity constraint), and (e) proposed entropic reconstruction in the domain defined by (c). Due to their very large dynamics, reconstructions in (d) and (e) are in logarithmic scale.

constraint, that is, the object exists is some areas and not in others, because

- they know *a priori* the position of the object;
- this introduces a kind of regularization, thus improving the conditioning of the problem and the quality of the reconstruction;
- this speeds up the reconstruction process.

Using the entropic frame, we can refine this approach, by using variable bounds on the intensity of the object to be reconstructed. These bounds are derived from the observations in Fig. 3 (a). This bounded domain $\bigotimes[a_i, b_i]$ is represented in Fig. 3 (c), with $a_i = 0$, $\forall i$ (positivity constraint). We then took a uniform measure over each interval, see Table I. The resulting criterion is not explicit and the problem was solved using its dual form

$$\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t \boldsymbol{z} - \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{A}^t\boldsymbol{\lambda}) - \mathcal{F}_{\boldsymbol{n}}^*(\boldsymbol{\lambda})$$

with

$$\begin{cases} \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{s}) = \displaystyle\sum_{j=1}^{K} \log\left(\frac{e^{a_j s_j} - e^{b_j s_j}}{s_j}\right) \\ \text{and} \quad \hat{\boldsymbol{x}} = \operatorname{grad} \mathcal{F}_{\boldsymbol{x}}^*(\boldsymbol{A}^t\hat{\boldsymbol{\lambda}}). \end{cases}$$

Due to their very large dynamics, the results of Fig. 3(d) and (e) are in logarithmic scale. They show the improvement that could be provided by refining the search space.

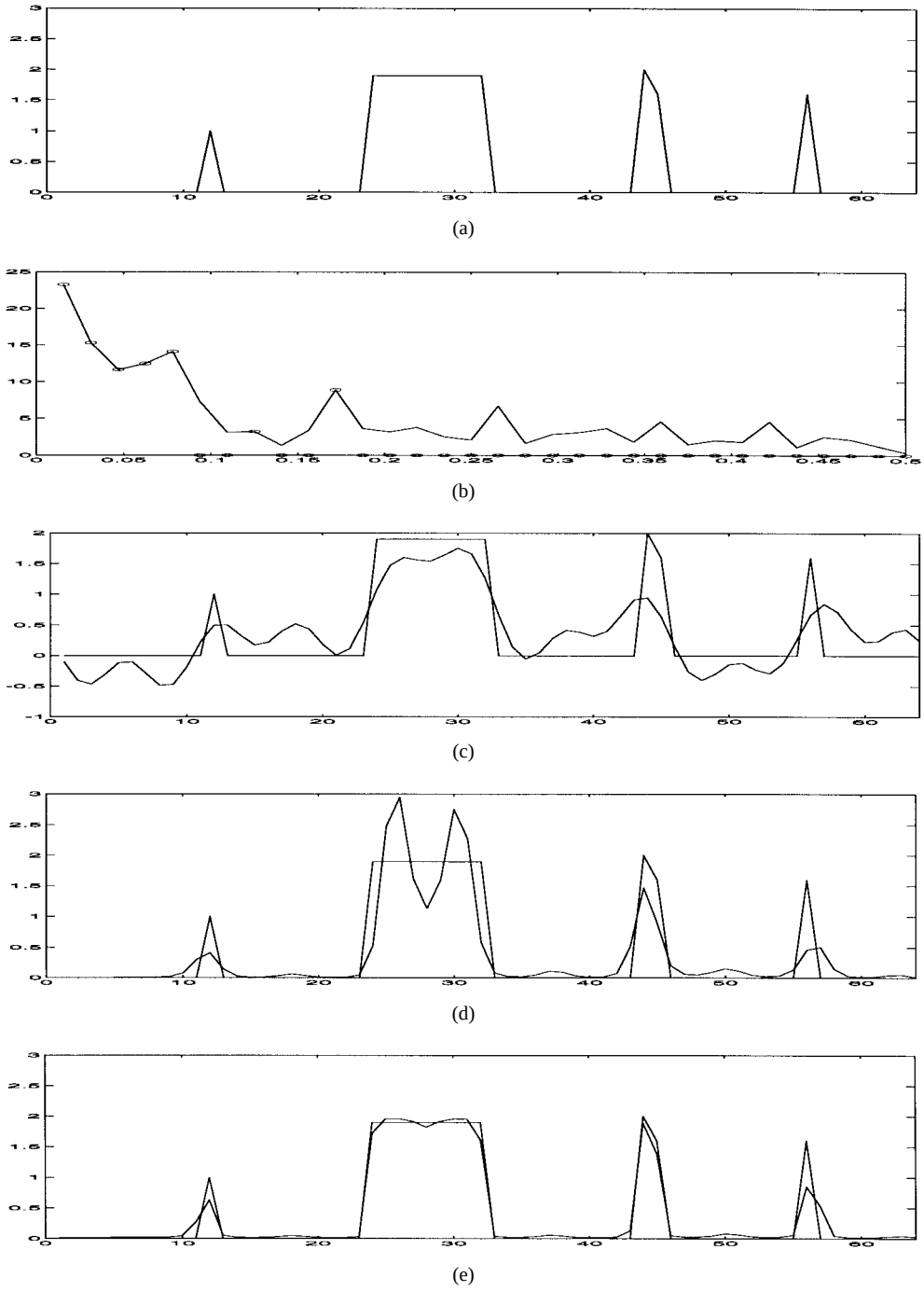


Fig. 4. Comparisons in Fourier synthesis: this figure compares different reconstructions in a simple Fourier synthesis problem. The test object is in (a), magnitude of its Fourier transform and of the available data in (b). Then three reconstructions corresponding to different reference measures of the proposed entropic scheme, and also to different constraint sets, are given. They show the improvement with the reduction of the set of admissible solutions.

2) Fourier Synthesis: Fourier synthesis is a classical inverse problem. Examples of such situations can be found in radioastronomy or tomography, see for instance [50], [51]. The following one-dimensional synthetic example, Fig. 4(a)–(e), is intended to illustrate the impact of the definition of the "set of admissible solutions," together with the flexibility of the method, which provides here three different criteria.

Fig. 4(a) shows the original signal made of 64 values between 0 and 2 (these values are given in Table II). The observation is accessible in the form of seven of the Fourier coefficients, for the points $\{1$ to $5,\ 8$ and $11\}$ of the discrete Fourier transform of the object, Fig. 4(b).

TABLE II
NONZERO VALUES OF THE TEST OBJECT $\boldsymbol{x}$

$x(12)$	$x(23)$ to $x(32)$	$x(44)$	$x(45)$	$x(56)$
1	1.9	2	1.6	1.6

As the operator is a Fourier matrix, the problem is well-conditioned and a nonregularized Fourier inversion yields a correct but low-resolution reconstruction shown in Fig. 4(c).

Taking positivity into account by using a Poisson reference measure (hence a "$x \log x$" regularizer) leads to the reconstruction of Fig. 4(d), which shows a far better resolution, but also a split of the main feature of the original signal, which cannot be removed without prejudicial decrease of resolution. Knowing that the object is actually between the bounds 0 and 2 can be accounted for with a reference measure made of product of uniform probability densities over $[0, 2]$, and after dual optimization of the implicit convex primal problem, a close-to-perfect reconstruction is obtained in Fig. 4 (e).

V. Conclusions

The general entropic framework we have presented is based on a combination of stochastic models designed to account for prior information both on the object and the measurement process—as in Bayesian estimation. The specific rule of combination of these "prior" measures is based on Kullback information. As such, the resulting criteria share interesting properties, some of them previously observed in axiomatic constructions of inversion processes, and others simplifying the practical implementation. All this is obtained with a degree of freedom provided by the choice of reference measures. Thus design of an efficient criterion in view of a particular applied problem is made possible by the proposed framework. New criteria have been presented and illustrated.

It should be mentioned that these results highly rely on convex analysis results. In particular, the importance of dual formulations is tantamount. For instance, some of the presented criteria are only defined by their dual formulation.

The entropic framework has been also linked with large deviation theory, see the work of F. Gamboa *et al.* for further developments on this subject.

Although we derived the methodology in view of a penalized approach of inversion, it should be emphasized that the resulting criteria are merely divergence measures between members of convex sets, and as such can be used in other contexts, for instance, set-theoretic estimation.

Among the features of this work is the explicit accounting of observation noise in the methodology. Indeed, this is an essential element in real inverse problems.

In view of real data processing, we have presented some practical guidelines and illustrations of the potential of the method. Along the same lines, the issue of enabling "correlations" between components of objects, and the case of a nonlinear operator have been considered.

Throughout this paper a general entropic framework for the resolution of linear inverse problems has been presented. It gives a unified view of numerous results and criteria for reconstruction—until now scattered in several fields (electrical engineering, physics, astronomy, crystallography, applied mathematics, and statistics)—and provides an "entropic" basis to many known reconstruction criteria.

Acknowledgment

The authors are indebted to the three anonymous referees for their careful reading of the manuscript and their numerous and useful comments that have improved the presentation of this paper.

References

[1] D. Mukherjee and D. C. Hurst, "Maximum entropy revisited," *Statistica Neerlandica*, vol. 38, no. 1, pp. 1–11, 1984.
[2] L. K. Jones and V. Trutzer, "Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method," *Inverse Problems*, vol. 5, pp. 749–766, 1989.
[3] L. K. Jones and C. L. Byrne, "General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 23–30, Jan. 1990.
[4] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26–37, Jan. 1980.
[5] I. Csiszár, "Why least-squares and maximum entropy—An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.
[6] ———, "Generalized projection for nonnegative functions," *Acta. Math. Hungar.*, vol. 6, nos. 1/2, pp. 161–185, 1995.
[7] J. Navaza, "On the maximum entropy estimate of electron density function," *Acta Crystallogr.*, pp. 232–244, 1985.
[8] ———, "The use of nonlocal constraints in maximum-entropy electron density reconstruction," *Acta Crystallogr.*, pp. 212–223, 1986.
[9] D. Dacunha-Castelle and F. Gamboa, "Maximum d'entropie et problème des moments," *Ann. l'Institut Henri Poincaré*, vol. 26, no. 4, pp. 567–596, 1990.
[10] F. Gamboa and É. Gassiat, "Bayesian methods and maximum entropy for ill posed inverse problems," *Ann. Statist.*, vol. 25, no. 1, Feb. 1997.
[11] ———, "The maximum entropy method on the mean: Applications to linear programming and superresolution," *Math. Programming*, no. 66, pp. 103–122, Oct. 1994.
[12] F. Gamboa and M. Lavielle, "On two-dimensional spectral realization," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1603–1608, Sept. 1994.
[13] I. Csiszár, F. Gamboa, and É. Gassiat, "MEM pixel correlated solutions for generalized moment and interpolation problems," *IEEE Trans. Inform. Theory*, submitted for publication, 1998.
[14] G. Le Besnerais, J. Navaza, and G. Demoment, "Aperture synthesis in astronomical radio-interferometry using maximum entropy on the mean," in *SPIE Conf., Stochastic and Neural Methods in Signal Processing, Image Processing and Computer Vision*, S. Chen, Ed. (San Diego, July 1991).
[15] J.-F. Bercher, G. Le Besnerais, and G. Demoment, "The maximum entropy on the mean method, noise and sensitivity," in *Maximum Entropy and Bayesian Methods*, J. Skilling, Ed. Dordrecht, The Netherlands: Kluwer, 1994.
[16] C. Heinrich, J.-F. Bercher, G. Le Besnerais, and G. Demoment, "Restoration of spiky signals: A new optimal estimate and a comparison," in *Proc. IEEE ICASSP* (Detroit, MI, May 1995), pp. 877–880.
[17] C. Heinrich, J.-F. Bercher, and G. Demoment, "The maximum entropy on the mean method, correlations and implementation issues," in *Workshop Maximum Entropy and Bayesian Methods.* Berg-en-Tal, South Africa: Kluwer, 1996, pp. 52–61.
[18] P. Maréchal and A. Lannes, "Unification of some deterministic and probabilistic methods for the solution of linear inverse problems via the principle of maximum entropy on the mean," *Inverse Probl.*, vol. 13, pp. 135–151, 1997.
[19] B. Urban, "Retrieval of atmospheric thermodynamical parameters using satellite measurements with a maximum entropy method," *Inverse Probl.*, vol. 12, pp. 779–796, 1996.
[20] C. E. Shannon, "The mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, pp. 623–656, 1948.
[21] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
[22] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959.
[23] I. Csiszár, "I—Divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.
[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.
[25] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkrantz, Ed., Reidel edition. Dordrecht, The Netherlands: Kluwer, 1982.
[26] B. Buck and V. A. Macaulay, Eds., *Maximum Entropy in Action.* Oxford, U.K.: Clarendon, 1991.

[27] M. A. Berger, *An Introduction to Probability and Stochastic Processes.* New York: Springer-Verlag, 1992.
[28] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics.* New York: Springer-Verlag, 1985.
[29] I. Csiszár, "An extended maximum entropy principle and a Bayesian justification," in *Bayesian Statistics 2*, J. M. Bernardo, M. H. DeGroot, and A. F. M. Smith, Eds. Amsterdam, The Netherlands: North-Holland Elsevier, 1985, pp. 83–98.
[30] R. Azencott, "Grandes déviations et applications," in *École D'été de Probabilité de Saint Flour VIII-1978*, P. L. Hennequin, Ed. Berlin, Germany: Springer-Verlag, 1978, pp. 2–172.
[31] A. Dembo and O. Zeitouni, *Large Deviations Techniques.* Boston, MA: Jones and Bartlett, 1992.
[32] I. Csiszàr and J. Körner, *Information Theory: Coding Theorems for Discrete Memory-Less Systems.* New York: Academic, 1981.
[33] C. Robert, "An entropy concentration theorem: Applications in artificial intelligence and descriptive statistics," *J. Appl. Probab.*, vol. 27, pp. 303–313, 1990.
[34] J. M. Borwein and A. S. Lewis, "Duality relationships for entropy-like minimization problems," *SIAM J. Contr. Optimiz.*, vol. 29, no. 2, pp. 325–338, 1991.
[35] A. Decarreau, D. Hilhorst, C. Lemaréchal, and J. Navaza, "Dual methods in entropy maximization. Application to some problems in crystallography," *SIAM J. Optimiz.*, vol. 2, no. 2, pp. 173–197, May 1992.
[36] O. Barndorff-Nielsen, *Information and Exponential Model in Statistics.* New York: Wiley, 1978.
[37] R. T. Rockafellar, *Convex Analysis.* Princeton, NJ: Princeton Univ. Press, 1970.
[38] I. Csiszár, "Sanov property, generalized I-projections and a conditional limit theorem," *Ann. Probab.*, vol. 12, pp. 768–793, 1984.
[39] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. and Math. Phys.*, vol. 7, pp. 200–217, 1967.
[40] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. and Commun.*, vol. 53-A, pp. 36–43, 1970.
[41] J. P. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Meet. Soc. Exploration Geophysicists* (Oklahoma City, OK, Oct. 1967), pp. 34–41.
[42] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 230–237, Apr. 1981.
[43] J. Skilling, "The axioms of maximum entropy," in *Maximum Entropy and Bayesian Methods*, G. J. Erickson and C. R. Smith, Eds. Dordrecht, The Netherlands: Kluwer, 1988, pp. 173–187.
[44] J. Shore and R. Johnson, "Which is the better entropy expression for speech processing: $-s\log s$ or $\log s$?," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, Feb. 1984.
[45] R. Narayan and R. Nityananda, "Maximum entropy image restoration in astronomy," *Ann. Rev. Astron. Astrophys.*, vol. 24, pp. 127–170, 1986.
[46] K. Horne, "Image of accretion disc—I: The eclipse mapping method," *Monthly Notes Roy. Astronom. Soc.*, vol. 213, pp. 129–141, 1985.
[47] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern, "Maximum entropy and the nearly black objects," *J. Roy. Statist. Soc. B*, vol. 54, no. 1, pp. 41–81, 1992.
[48] S. F. Gull, "Developments in maximum entropy data analysis," in *Maximum Entropy and Bayesian Methods*, J. Skilling, Ed. Dordrecht, The Netherlands: Kluwer, 1989, pp. 53–71.
[49] J. A. O'Sullivan, "Divergence penalty for image regularization," in *Proc. IEEE ICASSP* (Adelaide, Australia, Apr. 1994), pp. 541–544.
[50] G. T. Herman, H. K. Tuy, H. Langenberg, and P. C. Sabatier, *Basic Methods of Tomography and Inverse Problems.* Adams Hilger, 1987.
[51] A. R. Thomson, J. M. Moran, and G. W. Swenson, *Interferometry and Synthesis in Radioastronomy*. New York: Wiley, 1986.