MATHEMATICAL METHODS IN THE APPLIED SCIENCES
Math. Meth. Appl. Sci. 2007; **30**:1527–1544
Published online 21 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/mma.855
MOS subject classification: 47 A 52; 45 B 05; 94 A 17

Joint additive Kullback–Leibler residual minimization and regularization for linear inverse problems

Elena Resmerita[1, *, †] and Robert S. Anderssen[2]

[1] *Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria*
[2] *CSIRO Mathematical and Information Sciences, PO Box 664, Canberra, ACT 2601, Australia*

Communicated by R. P. Gilbert

SUMMARY

For the approximate solution of ill-posed inverse problems, the formulation of a regularization functional involves two separate decisions: the choice of the residual minimizer and the choice of the regularizor. In this paper, the Kullback–Leibler functional is used for both. The resulting regularization method can solve problems for which the operator and the observational data are positive along with the solution, as occur in many inverse problem applications. Here, existence, uniqueness, convergence and stability for the regularization approximations are established under quite natural regularity conditions. Convergence rates are obtained by using an *a priori* strategy. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: ill-posed problems; regularization; Kullback–Leibler distance; information; Radon–Nikodym theorem; Banach space adjoint operator

1. INTRODUCTION

We consider the following subclass of inverse problems:

$$Ax = y \tag{1}$$

where $A : X \to Y$ is a compact integral operator with a non-negative kernel a, with X and Y denoting Banach spaces. Such problems are ill posed in the sense of Hadamard [1], that is, the degree of the improper posedness increases with the smoothness of the kernel a. For their numerical solution, some form of regularization must be applied. In the classical approach, dating back to Tikhonov [2] and Twomey [3], the regularization functional $\mathscr{F}(x)$ is formulated as a weighted

[*] Correspondence to: Elena Resmerita, Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria.
[†] E-mail: elena.resmerita@ricam.oeaw.ac.at, elena.resmerita@oeaw.ac.at

Received 15 June 2006

sum of a residual minimizer and a regularizor

$$\mathscr{F}(x) = E[(Ax - y^{\delta})^2] + \alpha f(Tx), \quad 0<\alpha<\infty \tag{2}$$

where (*a*) E denotes a statistical expectation operator with $E[(Ax - y^{\delta})] = 0$, y^{δ} the noisy data, α the regularization (weight) parameter, and (*b*) the penalty function f and the differential operator $T : X \to X$ are chosen to mollify the inherent improper posedness encapsulated in the smoothness of the kernel a. Because the corresponding Euler–Lagrange equations are linear, a popular choice is to take both the residual minimizer and the regularizor to be quadratic functionals. The usual choice for T is the second derivative operator. For data smoothing, the corresponding unique regularization solution is a cubic spline [4].

With the residual minimizer a quadratic functional, non-quadratic choices for the regularizor such as Sobolev-type norms, semi-norms and entropy constructs (Boltzmann–Wiener–Shannon and Kullback–Leibler) have been examined extensively and proved successful from both theoretical and practical perspectives.

However, in many practical situations such as emission tomography [5, 6], where A is a positive operator, the solution x and the exact data y are probability distributions. This leads naturally to the idea that such a situation represents an opportunity where it might be appropriate to choose both the residual minimizer and the regularizor to be the same non-quadratic functional. Except for [7, 8], little appears to have been published that exploits this point. A natural choice is the functional proposed by Kullback and Leibler [9] (to be referred to as the *KL-functional*) which, for probability densities u and v, takes the form

$$d(u, v) = \int \left[u(t) \ln \frac{u(t)}{v(t)} \right] \mathrm{d}t \tag{3}$$

An alternative choice could be the symmetric KL-functional

$$J(u, v) = d(u, v) + d(v, u) = \int \left[(u(t) - v(t)) \ln \frac{u(t)}{v(t)} \right] \mathrm{d}t \tag{4}$$

When u and v are not densities, one can generalize the form of $d(u, v)$ by adding additional (linear) terms as long as they cancel if u and v were densities to recover the structure of (3). Among the various possibilities, we add, as detailed in Equation (10) of Section 3, the terms that transform $d(u, v)$ to become a Bregman distance (divergence). The analysis below specifically exploits the various properties of Bregman distances.

The KL-functional represents an interesting choice for the following reasons:

(i) As a residual minimizer, it corresponds to observed/measured Poisson errors, such as count errors in emission tomography. A detailed discussion of the statistical interpretation of $d(u, v)$ with respect to discrete observational data can be found in Section 2.1 of [10].

(ii) As a regularizor, it allows for the incorporation of prior information about the solution or some transformation of the solution.

It is this latter possibility that will be examined in this paper, where the regularization functional $\mathscr{F}_d(x)$ is chosen to have the form

$$\mathscr{F}_d(x) = d(y^{\delta}, Ax) + \alpha d(x, x^*) \tag{5}$$

DOI: 10.1002/mma

with x^* denoting a prior estimate of the solution and d being the Kullback–Leibler functional (10) defined for functions which are not necessarily densities. Now, however, the use of the functional $\mathscr{F}_d$ for the regularization of inverse problems imposes its own regularity constraints on the structure of problems that can be solved and the proofs that the regularization approximations converge and are stable. On the one hand, the KL-functionals are well defined only for inverse problems with non-negative solutions. On the other hand, in order to establish convergence and stability for the regularization approximations, we need to formulate conditions on the operator A and the solution x of Equation (1) which guarantee that the regularization approximations satisfy the same conditions as the solution x. This can be viewed as a separate step of proving that the regularization approximations have appropriate regularity. Consequently, from the perspective of regularization with non-quadratic functionals for both the residual minimizer and the regularizor, this is one of the key consideration in the construction of proofs for convergence and stability. Clearly, this is not an issue for the regularization of general inverse problems when quadratic functionals are chosen for the residual minimizer and the regularizor.

Though the joint use of non-quadratic residual minimizers and regularizors involves greater technical challenges in constructing proofs of convergence and stability, they do implicitly perform a type of regularization by limiting attention to a subset of functions in the (Banach) space on which the operator A is defined.

Under the assumption that the operators A, the solutions x and the perturbed data y^δ are positive and satisfy quite natural boundedness conditions, which hold for a very wide class of practical and industrial inverse problems, existence, uniqueness and convergence for the regularized approximations are derived below along with stability results and convergence rates.

The paper has been organized in the following manner. In Section 2, background is presented about properties of KL-functionals, their relationship to Bregman divergence and their earlier role as regularizors. In particular, the connection between information, statistical sufficiency and the Radon–Nikodym theorem is discussed in terms of the motivation behind the original formulation of the KL-functional as an information measure [9]. Section 3 focuses on the basic assumptions, notations and topological and algebraic properties of the Boltzmann–Shannon entropy and of the KL-functional. Existence, uniqueness and regularity properties of the regularized solutions are derived in Section 4, while stability and convergence results are established in Section 5. A discussion on the possibility of employing the symmetric KL-functional (4) as a regularizor completes Section 5. Convergence rates are derived in Section 6 under a suitable sourcewise representation of the true solution.

2. KULLBACK–LEIBLER FUNCTIONAL

Within the statistical and information theoretic literature, the KL-functional and its symmetric form (4) are popular because they arise as the consequence of quite independent considerations. The work reported in [11] appears to have been the first to point out that the symmetric form (4) had special statistical properties from a Bayesian perspective. In particular, [11] established that (4) was one of a limited number of 'non-informative' ('invariant' in the paper's terminology) priors in the sense that it contains the same information irrespective of its position on the real line. It was Kullback and Leibler [9] who proved that the functionals $d(u, v)$, $d(v, u)$ and $J(u, v)$ were invariant (in their sense) when used to define 'information'. As explained below, in order to prove their invariance, the authors of [9] utilized the measure theoretic definition of 'statistical

DOI: 10.1002/mma

sufficiency' from [12], for which the Radon–Nikodym theorem plays a central role. As noted in [12], this theorem is 'essential to the very definition of the basic concept of conditional probability' for a subset of a sample space given the value of a sufficient statistic.

The 'self-consistency' (in an inductive inference sense) of cross-entropy (an alternative name for the KL-functional) was established in [13], together with Jaynes' principle of maximum entropy [14], with respect to accommodating new information given as expected values. The book [15] examines the properties of cross-entropy in considerable detail by, among other things, exploiting its expectation interpretation. It is noted there that the cross-entropy is a special case of the Ali–Silvey 'distance' [16] and its importance from a Bayesian perspective [17] is also discussed. Two recent papers [18, 19] derived estimates for the approximation of the entropy function $x \ln x$ and examined the minimization of the KL-functional for unknown finite distributions, respectively.

2.1. Motivation: information and sufficiency

As detailed in [9], the concepts of information and statistical sufficiency are intimately interconnected. To formalize the mathematical foundations of theoretical statistics, the criterion of sufficiency was introduced in [20, p. 316], requiring that a chosen statistic 'should summarize the whole of the relevant information supplied by the sample'. Subsequently, the concept of a sufficient statistic was defined in [21] as being 'equivalent, for all subsequent purposes of estimation, to the original data from which it was derived'. Thus, formally, a sufficient statistic is a statistic for which the conditional distribution of the original data, given that statistic, depends only on that data. This leads naturally to the idea that the values of any definition of information must, with respect to the set of sufficient statistics transformations that can be applied to the given measurements, be invariant over the corresponding set of transformed measurements.

As established in [9], the importance of the functional (3) is that it is invariant (not decreased) if and only if sufficient statistics transformations are applied to the given measurements. In those deliberations, the earlier results in [12] play a central role. From a regularization perspective, the importance of this result is that, by using the KL-functional, one is specifically examining, with respect to a set of given measurements, the application of all possible choices of sufficient statistics, rather than a specific choice such as the minimum variance functional. Scientifically, the success of the KL-functional relates to the fact that it has numerous properties that can be exploited mathematically. A detailed discussion of such results can be found in [22].

2.2. The KL-functional as a regularizor

More recently, the role of the KL-functional in the solution of inverse problems has received considerable attention. As explained in the Introduction, it is this aspect that is the focus of this paper. A quite theoretical paper [8], which treats regularization more from an optimization rather than a variational perspective, derived detailed results about a generalized entropic regularization framework which includes $\mathscr{F}_d$ regularization, as formulated above, as a special case. The Radon–Nikodym theorem plays a central role there. The issues addressed in that paper are complementary and supplementary to the results derived below. For example, [8, p. 1570] establishes a link between entropic functions and Bregman divergences [23]. In an examination of an axiomatic approach to inference for linear inverse problems, [24] compares the roles of least squares and entropy measures.

From a practical perspective, because of its minimum information interpretation, entropy regularization has been popularized by various authors. Jaynes [25] proposed it as an appropriate way for solving indirect measurement problems, while Amato and Hughes [26] and Engl and Landl [27] showed convergence and Eggermont [28] established convergence rates. In [29], its regularization performance was analysed in terms of statistical shrinkage concepts. In hindsight, as noted by [30], its real appeal is related to the fact that it guarantees positivity for the regularized solution when the KL-functional is used as the regularizor. More recently, it has found applications in machine learning research [31].

2.3. KL-functional perturbations and convergence

As explained above, the KL-functional allows one to perform regularization from an information perspective, in the sense that one constrains, for example, the closeness of the data y to its observed perturbation y^{δ} to satisfy an information measure rather than some distance measure associated with some function space. In this way, a more natural characterization of the relationship between the data and its observational realization is achieved, and thereby becomes a basis for assessing the convergence of the regularized solution to the exact.

Furthermore, this does not compromise the analysis as, under similar regularity to that invoked to prove convergence in the KL-framework, related strong convergence results can be established (e.g. the Proposition in Section 5.2).

Within a regularization context, the motivation for the condition $d(y^{\delta}, y)\leqslant\delta^2$ used in Section 5 corresponds to the natural asymmetric discrimination interpretation of $d(y^{\delta}, y)$: when y^{δ} and y correspond to probability density distributions F and G, respectively, the quantity $d(y^{\delta}, y)$ can be viewed as a measure of the amount of information in a single data value from the distribution of F for discrimination between F and G.

The role and appropriateness of KL-divergence as a statistical measure have been investigated in some detail in [10, 32, 33].

2.4. The Bregman divergence connection

As noted in [34], the key ingredient in [28] for the analysis of entropy regularization for first kind Fredholm integral equations is the Bregman divergence. Its properties and associated inequalities, as shown in [28], are essential when analysing convergence of the method and establishing convergence rates. Similar inequalities were derived in [35, 36]. In fact, some of the methodology developed in [28] plays a role at some of the stages in the analysis below.

Among others, Bregman divergence becomes, for appropriate choices of a generating functional, the Hilbert space norm and the KL-functional. The generality of this framework for the analysis of entropy regularization and its extensions has proved successful in establishing error estimates and convergence rates [34, 37]. A detailed discussion about its background and relationship to regularization can be found in [34].

2.5. Relationship to penalized maximum likelihood

There are different ways in which penalized maximum likelihood can be motivated and formulated. The paper [38] takes the direct approach by defining it as an additive regularization formalism and then reinterpreting it from a statistical perspective using conditional and prior probability considerations. In [39], the formulation is based on Bayes' theorem and the

DOI: 10.1002/mma

reinterpretation of the conditional probability distribution as its likelihood dual. The classical additive regularization structure is then derived on taking the logarithm of the Bayesian multiplicative formula and reinterpreting the logarithm of the prior distribution to be the weighted regularizor. The type of regularization formulation investigated in [7, 8, 34, 37] is generated on reinterpreting the conditional probability to be an information measure, such as the KL-functional, instead of a likelihood.

Historically, Good and Gaskins [40] appear to be the first to appreciate conceptually the need for regularization in the solution of statistical problems. Their motivation and terminology has a strong statistical and data smoothing, rather than a mathematical, emphasis, with a natural connection to non-parametric methods which can be seen as a precursor to spline [4] and kernel method techniques [41]. The subsequent identification and exploitation by various authors of this duality between penalized likelihood and regularization has resulted in a synergetic exchange of ideas and techniques between statistics and (applied) mathematics. In turn, this has allowed the challenge of constructing stable algorithms for observational data for the recovery of information from indirect measurement problems to be successfully accommodated. A representative illustration is the work by Eggermont and LaRiccia [42]. The regularization studied there is constructed to have the KL-functional as the residual minimizer and the Good and Gaskins roughness penalty [40] as the regularizor

$$\min_{x\geqslant 0}\{d(y, Ax) + \alpha\|\nabla\sqrt{x}\|_2^2\} \tag{6}$$

Among other things, it was shown that this form of regularization is intimately connected to the EM (Estimate and Maximize) algorithm [43] for maximum likelihood functional estimation. In [42], the special structure of the regularization functional is successfully exploited to establish some quite specific results including

$$\lim_{\alpha\to 0}\|\sqrt{x_\alpha} - \sqrt{\bar{x}}\|_{H^1} = 0$$
$$\lim_{\alpha\to 0}\|x_\alpha - \bar{x}\|_C = 0$$

where x_α denotes the exact regularization solution, and $\bar{x}$ is the solution of the problem

$$\min_{x\geqslant 0}\|\nabla\sqrt{x}\|_2^2 \quad \text{subject to } Ax = y \tag{7}$$

Rates of convergence in L^1 are also established for a specified source condition.

Given that Dempster *et al.* [44] formulated the EM methodology for maximizing the likelihood of incomplete data, it is not surprising that this methodology has an algorithmic connection to penalized maximum likelihood. However, because of the known poor performance of EM when applied directly to the recovery of information from indirect measurement problems [39], it seems natural to turn to the direct solution of an appropriate penalized maximum likelihood counterpart. This is the approach taken in [45] for the stabilized recovery of information for additive Poisson regression problems, and in [46] for the construction of two-stage-splitting algorithms for the maximization of matrix versions of penalized likelihood formulations.

3. NOTATIONS, ASSUMPTIONS AND PRELIMINARY RESULTS

3.1. The Boltzmann–Shannon entropy is the function $g : L^1(\Omega) \to (-\infty, +\infty]$ with $\Omega \in \mathbb{R}^n$ bounded and measurable, given by[‡]

$$g(x) = \begin{cases} \displaystyle\int_\Omega x(t) \ln x(t)\, \mathrm{d}t & \text{if } x \geqslant 0 \text{ a.e. and } x \ln x \in L^1(\Omega) \\ +\infty & \text{otherwise} \end{cases} \tag{8}$$

The Kullback–Leibler functional (denoted below by d) or the Bregman divergence (distance) with respect to the Bolzmann–Shannon entropy can be defined for functions which are not necessarily probability densities. More precisely, one defines $d : \operatorname{dom} g \times \operatorname{dom} g \to [0, +\infty]$ by

$$d(v, u) = g(v) - g(u) - g'(u, v - u) \tag{9}$$

where $g'(u, \cdot)$ is the directional derivative of g at u. One can also write

$$d(v, u) = \int_\Omega \left[v(t) \ln \frac{v(t)}{u(t)} - v(t) + u(t) \right] \mathrm{d}t \tag{10}$$

when $d(v, u)$ is finite.

3.2. As motivated in the Introduction, we would like to approximate solutions of Equation (1) *via* the following auxiliary problem:

$$\min_{x \geqslant 0} d(y^\delta, Ax) + \alpha d(x, x^*) \tag{11}$$

where x^* contains *a priori* information and y^δ denote the perturbed data.

We assume the following:

(A1) The operator $A : L^1(\Omega) \to L^1(\Sigma)$ is linear and compact.

(A2) The operator A satisfies $Ax > 0$, a.e. for any $x > 0$ a.e.

(A3) The maximum entropy solution $\bar{x}$ of Equation (1) exists, that is, there exists a positive function $\bar{x} \in L^1(\Omega)$ which is the solution of the problem

$$\begin{aligned} \min \quad & d(x, x^*) \\ \text{subject to} \quad & Ax = y \end{aligned} \tag{12}$$

(A4) The function x^* is bounded and bounded away from zero, i.e. there exist $k_1, k_2 > 0$ such that $k_1 \leqslant x^* \leqslant k_2$, almost everywhere on Ω.

(A5) For any $\delta > 0$, the data y^δ are bounded and bounded away from zero, almost everywhere on Σ.

(A6) If $x \in L^1(\Omega)$ is such that $c_1 \leqslant x \leqslant c_2$, a.e. for some positive constants c_1, c_2, then there exist $c_3, c_4 > 0$ such that $c_3 \leqslant Ax \leqslant c_4$, almost everywhere on Σ.

[‡] We use the convention $0 \ln 0 = 0$.

DOI: 10.1002/mma

Remarks
In most practical situations, when a reasonable level of data has been collected, the measured data will be non-zero over its domain of definition. For example, in emission tomography, after a suitably long sampling time, all counters will have recorded some photon activity, even if very small. Consequently, assumption (A5) holds for such situations. Furthermore, assumption (A5) does not stop y^δ from having high-frequency components. Assuming that the theoretical structure of the inverse problem has a uniquely defined solution, it is the sensitivity of the inversion to high-frequency perturbations in the data that generates the need for regularization.

Assumption (A6) is not significantly restrictive. There are large classes of linear (even severely) ill-posed integral equations for which such a requirement is achieved. For example, one could think of integral equations of the first kind that arise in geophysics and potential theory, which have smooth, bounded and bounded away from zero kernels.

3.3. Recall below a lemma from [34], that will be needed in the sequel.

Lemma
The function defined by (8) has the following properties:

(i) The domain of the function g is strictly included in $L^1_+(\Omega) = \{x \in L^1(\Omega) : x \geqslant 0 \; a.e.\}$.
(ii) The interior of the domain of the function g with respect to the $L^1(\Omega)$ norm topology is empty.
(iii) The set $\partial g(x)$ is non-empty if and only if x belongs to $L^\infty_+(\Omega)$ and is bounded away from zero. Moreover, $\partial g(x) = \{1 + \ln x\}$.
(iv) The directional derivative of the function g is given by

$$g^\circ(x, v) = \int_\Omega v(t)[1 + \ln x(t)] \, \mathrm{d}t$$

whenever it is finite.
(v) For any $x, y \in \operatorname{dom} g$, one has

$$\|y - x\|_1^2 \leqslant \left(\tfrac{2}{3}\|y\|_1 + \tfrac{4}{3}\|x\|_1\right) d(y, x) \tag{13}$$

Corollary
If $\{u_\lambda\}_\lambda$, $\{v_\lambda\}_\lambda$ are sequences in $L^1(\Omega)$ such that one of them is bounded, then

$$\lim_\lambda d(v_\lambda, u_\lambda) = 0 \implies \lim_\lambda \|v_\lambda - u_\lambda\|_1 = 0 \tag{14}$$

Proof
One applies (v) stated above. □

3.4. The next lemma is a collection of basic results about entropy and the Kullback–Leibler distance, which will be repeatedly used in this paper.

Lemma
The following statements hold:

(i) The function $(v, u) \mapsto d(v, u)$ is convex and thus, so is the function $(v, x) \mapsto d(v, Ax)$.
(ii) The function $d(\cdot, x^*)$ is lower semicontinuous with respect to the weak topology of $L^1(\Omega)$.

DOI: 10.1002/mma

(iii) For any fixed $v \in \operatorname{dom} g$, the function $x \mapsto d(v, Ax)$ is lower semicontinuous with respect to the weak topology of $L^1(\Omega)$.

(iv) For any $C>0$ and any non-negative $u \in L^1(\Omega)$, the following sets are weakly compact in $L^1(\Omega)$:

$$\{x \in L^1(\Omega) : d(x, u) \leqslant C\}$$

Proof

(i) See [47]. (ii) See [28, Corollary 2.2]. (iv) See [28, Lemma 2.1]. (iii) The proof is similar to the one for (ii). For the sake of completeness, we detail it below. Fix $v \in \operatorname{dom} g$. Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in the domain of the function $z \mapsto d(v, Az)$, which converges in the $L^1(\Omega)$ norm to some $x \in L^1_+(\Omega)$. Then, it converges to x almost everywhere on Ω. Also, continuity of the operator A yields convergence of $\{Ax_n\}_{n\in\mathbb{N}}$ to Ax in $L^1(\Sigma)$, as well as convergence almost everywhere. Consequently, the sequence $v \ln(v/Ax_n) - v + Ax_n$ converges almost everywhere to $v \ln(v/Ax) - v + Ax$. One can apply Fatou's Lemma and conclude

$$\int_\Omega [v \ln(v/Ax) - v + Ax] \, \mathrm{d}\mu \leqslant \liminf_{n\to\infty} \int_\Omega [v \ln(v/Ax_n) - v + Ax_n] \, \mathrm{d}\mu$$

which means that the function $d(y, A\cdot)$ is lower semicontinuous. Since it is also convex (cf. (i)), its weak lower semicontinuity is guaranteed. □

4. EXISTENCE, UNIQUENESS AND REGULARITY OF APPROXIMANTS

4.1. In this section, it is shown that problem (11) has a unique solution which, within the current theoretical framework, satisfies a specific regularity property. We begin with proving consistency of the regularized problem.

Proposition

For any $\alpha>0$ and $\delta>0$, there exists a unique solution x_α^δ of problem (11).

Proof

In order to prove existence of solutions, one shows that the function $d(y^\delta, A\cdot) + \alpha d(\cdot, x^*)$ is weakly lower semicontinuous on $L^1(\Omega)$ and that for any $C>0$, the sublevel sets

$$\{x \in L^1(\Omega) : d(y^\delta, Ax) + \alpha d(x, x^*) \leqslant C\}$$

are weakly compact in $L^1(\Omega)$. The weak lower semicontinuity property is a consequence of Lemma 3.4(ii)–(iii). Since the above sets are weakly closed subsets of the weakly compact sublevel sets of the function $d(\cdot, x^*)$, as stated in Lemma 3.4(iv), it follows that they are weakly compact, too. Therefore, there exist solutions x_α^δ of problem (11). Moreover, the solution is unique, due to the strict convexity of $d(\cdot, x^*)$ (see, e.g. [48]). □

4.2. Boundedness and boundedness away from zero of the regularized solutions are established below.

Proposition

For any fixed $\alpha>0$ and $\delta>0$, the solution x_α^δ of problem (11) has the following property.

DOI: 10.1002/mma

There exist positive constants c_1, c_2 such that $c_1 \leqslant x_\alpha^\delta / x^* \leqslant c_2$, almost everywhere on Ω. Moreover, x_α^δ is bounded and bounded away from zero almost everywhere.

Proof
Claim. There exists a constant $c_1 > 0$ *such that* $x_\alpha^\delta / x^* \geqslant c_1$, *a.e. on* Ω. In order to prove this, we adapt an idea from [28] to our context. Suppose by contradiction that, for any $\varepsilon > 0$, there is a set U_ε with positive Lebesgue measure $\mu(U_\varepsilon)$, such that

$$\frac{x_\alpha^\delta(t)}{x^*(t)} < \varepsilon, \quad \forall t \in U_\varepsilon \tag{15}$$

Denote by χ_ε the characteristic function of the set U_ε and by h the function

$$h(\rho) = d(y^\delta, Ax_\alpha^\delta + \rho A \chi_\varepsilon) + \alpha d(x_\alpha^\delta + \rho \chi_\varepsilon, x^*), \quad \rho \geqslant 0$$

Observe that the convex function h has $\rho = 0$ as a minimizer. Consequently, the first order necessary optimality condition yields

$$h'(0, \rho) \geqslant 0, \quad \forall \rho \geqslant 0 \tag{16}$$

In order to establish the formula for $h'(0, \rho)$, one can use the expression of the directional derivative of the functional $x \mapsto d(y^\delta, Ax)$ which has already been determined in [42]:

$$d(y^\delta, A\cdot)'(x, u) = \int_\Sigma \left[-y^\delta(s) \frac{Au(s)}{Ax(s)} + Au(s) \right] \mathrm{d}s \tag{17}$$

It can be shown, by using the Monotone Convergence Theorem, that the directional derivative of $x \mapsto d(x, x^*)$ has the following form:

$$d(\cdot, x^*)'(x, u) = \int_\Omega u(t) \ln \frac{x(t)}{x^*(t)} \, \mathrm{d}t$$

Hence, inequality (16) becomes

$$\rho \int_\Sigma \left[-y^\delta(s) \frac{A\chi_\varepsilon(s)}{Ax_\alpha^\delta(s)} + A\chi_\varepsilon(s) \right] \mathrm{d}s + \alpha\rho \int_\Omega \chi_\varepsilon(t) \ln \frac{x_\alpha^\delta(t)}{x^*(t)} \, \mathrm{d}t \geqslant 0$$

for any $\rho \geqslant 0$. This implies

$$\alpha \int_{U_\varepsilon} \ln \frac{x_\alpha^\delta(t)}{x^*(t)} \, \mathrm{d}t + \int_\Sigma A\chi_\varepsilon(s) \, \mathrm{d}s \geqslant \int_\Sigma y^\delta(s) \frac{A\chi_\varepsilon(s)}{Ax_\alpha^\delta(s)} \, \mathrm{d}s \tag{18}$$

By Fubini's Theorem, it follows that

$$\begin{aligned} \int_\Sigma A\chi_\varepsilon(s) \, \mathrm{d}s &= \int_\Sigma \int_\Omega a(s,t) \chi_\varepsilon(t) \, \mathrm{d}t \, \mathrm{d}s \\ &= \int_\Omega \left(\int_\Sigma a(s,t) \, \mathrm{d}s \right) \chi_\varepsilon(t) \, \mathrm{d}t \end{aligned}$$

DOI: 10.1002/mma

$$
\begin{aligned}
&= \int_{U_\varepsilon} \left(\int_\Sigma a(s,t)\,\mathrm{d}s \right) \mathrm{d}t \\
&= \int_{U_\varepsilon} A^*\mathbf{1}(t)\,\mathrm{d}t \\
&\leqslant \|A^*\mathbf{1}\|_\infty \mu(U_\varepsilon)
\end{aligned}
$$

where $A^*\mathbf{1}$ denotes the adjoint operator from $L^\infty(\Sigma)$ to $L^\infty(\Omega)$ applied to the function which is almost everywhere equal to 1. By combining the last inequality with (18) and (15), one gets

$$
\mu(U_\varepsilon)(\alpha \ln \varepsilon + \|A^*\mathbf{1}\|_\infty) \geqslant \int_\Sigma y^\delta(s) \frac{A\chi_\varepsilon(s)}{Ax_\alpha^\delta(s)}\,\mathrm{d}s \tag{19}
$$

Then, one can choose ε small enough to obtain $\alpha \ln \varepsilon + \|A^*\mathbf{1}\|_\infty < 0$, which yields a contradiction for (19) because the term on the right-hand side there is non-negative. Thus, the proof of the claim is completed.

It remains to show that x_α^δ/x^* is bounded from above almost everywhere. To this end, suppose that the contrary holds. That is, for any positive ε there exists a set V_ε with positive measure $\mu(V_\varepsilon)$ such that

$$
\frac{x_\alpha^\delta(t)}{x^*(t)} > \varepsilon, \quad \forall t \in V_\varepsilon \tag{20}
$$

Let $cV_\varepsilon = \Omega \setminus V_\varepsilon$. One distinguishes two situations. First, if $\mu(cV_\varepsilon) = 0$, it follows that (20) holds a.e. on Ω. Since ε was arbitrarily chosen, one obtains a contradiction. Second, if $\mu(cV_\varepsilon) > 0$, then inequality (15) holds on cV_ε. Hence, the reasoning done above for U_ε, starting after formula (15) and ending at (19), applies this time to cV_ε. As a consequence, one reaches a contradiction.

Since x^* is also minorized and majorized by two positive constants (cf. (A4)), it follows that the solution x_α^δ has the same property. □

Corollary
For any fixed $\alpha > 0$ and $\delta > 0$, the function Ax_α^δ is minorized and majorized by two positive constants, almost everywhere on Σ.

Proof
The corollary is an immediate consequence of assumption (A6) and of the previous proposition. □

Remark
The last result, as well as hypotheses (A4)–(A5), guarantees not only that Ax_α^δ, y^δ and x^* have finite L^∞ norm, but also that $\ln Ax_\alpha^\delta$, $\ln y^\delta$ and $\ln x^*$ belong to L^∞. This will be needed in the subsequent analysis.

DOI: 10.1002/mma

5. STABILITY AND CONVERGENCE OF THE METHOD

5.1. Stability of the regularized problem (11) with respect to a certain kind of data perturbations is shown below. In many practical situations, it is not necessarily the data y^δ for which the solution of problem (11) is constructed, but some discretization of it. Consequently, a stability result is required that guarantees that the regularized approximations converge to the exact solution as the discretized data converges to the exact data. For the current regularization framework being investigated, it is natural to assess the convergence in terms of the KL-functional, as detailed below in equality (21). Consequently, from a discrimination perspective, the information in y_n must approach that in y^δ as n goes to infinity.

Proposition
Fix $\alpha>0$ and $\delta>0$. Suppose that y_n, $n\in\mathbb{N}$, are approximations in $L^1(\Sigma)$ of y^δ in the following sense:

$$\lim_{n\to\infty} d(y_n, y^\delta)=0 \tag{21}$$

Then, the sequence of solutions x_n for problem (11) corresponding to data y_n converges in the L^1-norm to the solution x_α^δ of the regularized problem corresponding to data y^δ.

Proof
Fix $\alpha>0$. For any $n\in\mathbb{N}$, the definition of x_n implies

$$d(y_n, Ax_n)+\alpha d(x_n, x^*)\leqslant d(y_n, Ax_\alpha^\delta)+\alpha d(x_\alpha^\delta, x^*) \tag{22}$$

One can show that the sequence $\{d(y_n, Ax_\alpha^\delta)\}_{n\in\mathbb{N}}$ is bounded. To this end, observe that the sequence $\{y_n\}_{n\in\mathbb{N}}$ converges strongly to y^δ in L^1 as well as pointwise almost everywhere, because $\lim_{n\to\infty} d(y_n, y^\delta)=0$ (see Corollary 3.3). Also, one has (see Remark in subsection 4.2)

$$\begin{aligned}|d(y_n, Ax_\alpha^\delta)-d(y^\delta, Ax_\alpha^\delta)-d(y_n, y^\delta)| &= \left|\int_\Omega (\ln Ax_\alpha^\delta-\ln y^\delta)(y_n-y^\delta)\,\mathrm{d}\mu\right| \\ &\leqslant \|\ln Ax_\alpha^\delta-\ln y^\delta\|_\infty \|y_n-y^\delta\|_1\end{aligned}$$

implying

$$\lim_{n\to\infty} d(y_n, Ax_\alpha^\delta)=d(y^\delta, Ax_\alpha^\delta) \tag{23}$$

Since the sequence $\{d(y_n, Ax_\alpha^\delta)\}_{n\in\mathbb{N}}$ is convergent, it is also bounded. By this fact together with (22), one gets boundedness of the sequence $\{d(x_n, x^*)\}_{n\in\mathbb{N}}$. Then, Lemma 3.4(iv) ensures existence of a subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ of $\{x_n\}_{n\in\mathbb{N}}$, which converges weakly to some $u\in L^1(\Omega)$. Actually, the element u lies in $\operatorname{dom} g$, because

$$d(u, x^*)\leqslant \liminf_{k\to\infty} d(x_{n_k}, x^*)<\infty$$

Compactness of the operator A implies strong convergence of the sequence $\{Ax_{n_k}\}_{k\in\mathbb{N}}$ to Au in $L^1(\Sigma)$ and hence, pointwise almost everywhere convergence. Then, Fatou's Lemma can be applied

DOI: 10.1002/mma

to the sequence $\{y_{n_k}\ln(y_{n_k}/Ax_{n_k}) - y_{n_k} + Ax_{n_k}\}_{k\in\mathbb{N}}$ and yields

$$d(y^\delta, Au) \leqslant \liminf_{k\to\infty} d(y_{n_k}, Ax_{n_k}) \tag{24}$$

Due to the weak lower semicontinuity of the function $d(\cdot, x^*)$ and due to (22) and (24), one also has

$$\begin{aligned} d(y^\delta, Au) + \alpha d(u, x^*) &\leqslant \liminf_{k\to\infty} d(y_{n_k}, Ax_{n_k}) + \alpha \liminf_{k\to\infty} d(x_{n_k}, x^*) \\ &\leqslant \liminf_{k\to\infty} [d(y_{n_k}, Ax_{n_k}) + \alpha d(x_{n_k}, x^*)] \\ &\leqslant \limsup_{k\to\infty} [d(y_{n_k}, Ax_{n_k}) + \alpha d(x_{n_k}, x^*)] \\ &\leqslant \limsup_{k\to\infty} [d(y_{n_k}, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*)] \\ &= d(y^\delta, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*) \end{aligned}$$

This means that u is the unique minimizer of problem (11), that is, $u = x_\alpha^\delta$. Consequently, it also follows that

$$d(y^\delta, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*) = \lim_{k\to\infty} [d(y_{n_k}, Ax_{n_k}) + \alpha d(x_{n_k}, x^*)] \tag{25}$$

In order to prove strong convergence of $\{x_{n_k}\}_{k\in\mathbb{N}}$ to x_α^δ, it is enough showing convergence with respect to the entropy g (see [48, Lemma 2.5]), that is, $\lim_{k\to\infty} g(x_{n_k}) = g(x_\alpha^\delta)$ or, equivalently, $\lim_{k\to\infty} d(x_{n_k}, x^*) = d(x_\alpha^\delta, x^*)$. Suppose, by contradiction, that

$$l = \limsup_{k\to\infty} d(x_{n_k}, x^*) > \liminf_{k\to\infty} d(x_{n_k}, x^*) \tag{26}$$

Let $\{x_j\}_{j\in\mathbb{N}}$ denote a subsequence of $\{x_{n_k}\}_{k\in\mathbb{N}}$ such that $l = \lim_{j\to\infty} d(x_j, x^*)$. By using (25), one gets

$$d(y^\delta, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*) = \lim_{j\to\infty} [d(y_j, Ax_j) + \alpha d(x_j, x^*)]$$

Consequently, by combining (25) and (26), one obtains

$$\begin{aligned} \lim_{j\to\infty} d(y_j, Ax_j) &= d(y^\delta, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*) - \alpha l \\ &< d(y^\delta, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*) - \alpha \liminf_{j\to\infty} d(x_j, x^*) \\ &\leqslant d(y^\delta, Ax_\alpha^\delta) \end{aligned}$$

The last inequality, which is due to the weak lower semicontinuity of the function $d(\cdot, x^*)$, is in contradiction with (24) since $u = x_\alpha^\delta$. Therefore, the sequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ converges in the $L^1(\Omega)$ norm to x_α^δ. Since, in fact, every convergent subsequence of $\{x_n\}_{n\in\mathbb{N}}$ converges strongly

DOI: 10.1002/mma

to the unique minimizer x_α^δ of problem (11), it follows that the whole sequence has the strong limit x_α^δ. □

5.2. The following result establishes convergence of the regularization method in the presence of 'entropy-perturbed' data, i.e. when y^δ satisfies

$$d(y^\delta, y) \leqslant \delta^2 \tag{27}$$

with $\delta > 0$. Condition (27) is a constraint on the possible perturbations of y to give a y^δ. It implies that, with respect to the KL measure of divergence, the admissible perturbations are such that the information in y^δ remain close to that in y.

Proposition
If the noisy data y^δ satisfy inequality (27), then the regularized solutions x_α^δ converge strongly to the maximum entropy solution $\bar{x}$ of Equation (1) as soon as $\delta \to 0$, $\alpha \to 0$ with $\delta^2/\alpha \to 0$.

Proof
According to the definition of x_α^δ, one gets

$$\begin{aligned} d(y^\delta, Ax_\alpha^\delta) + \alpha d(x_\alpha^\delta, x^*) &\leqslant d(y^\delta, A\bar{x}) + \alpha d(\bar{x}, x^*) \\ &\leqslant \delta^2 + \alpha d(\bar{x}, x^*) \end{aligned} \tag{28}$$

Consider α_n, δ_n such that $\alpha_n \to 0$ and $\delta_n^2/\alpha_n \to 0$. Let $\{x_n\}_{n\in\mathbb{N}}$ denote the sequence with term $x_{\alpha_n}^{\delta_n}$. The second inequality in (28) yields

$$d(x_n, x^*) \leqslant \frac{\delta_n^2}{\alpha_n} + d(\bar{x}, x^*) \tag{29}$$

Since the sublevel sets of the function $d(\cdot, x^*)$ are weakly compact (cf. Lemma 3.4(iv)), it follows that $\{x_n\}_{n\in\mathbb{N}}$ is contained in such a set. Then, there is a subsequence of it, denoted $\{x_{n_k}\}_{k\in\mathbb{N}}$, which converges weakly to some $v \in \operatorname{dom} g$. Hence, by the weak lower semicontinuity of the function $d(\cdot, x^*)$ and by (29), one obtains

$$d(v, x^*) \leqslant \liminf_{k\to\infty} d(x_{n_k}, x^*) \leqslant \limsup_{k\to\infty} d(x_{n_k}, x^*) \leqslant d(\bar{x}, x^*) \tag{30}$$

Due to (27) and Corollary 3.3, one has $\lim_{k\to\infty} \|y^{\delta_{n_k}} - y\|_1 = 0$. On one hand, (28) yields

$$\lim_{k\to\infty} d(y^{\delta_{n_k}}, Ax_{n_k}) = 0$$

By Corollary 3.3, this further implies that $\lim_{k\to\infty} \|y^{\delta_{n_k}} - Ax_{n_k}\|_1 = 0$. Consequently, $\lim_{k\to\infty} \|Ax_{n_k} - y\|_1 = 0$. On the other hand, compactness of the operator A implies that $\lim_{k\to\infty} \|Ax_{n_k} - Av\|_1 = 0$ and then $Av = y$, i.e. v is a solution of Equation (1). This combined with (30) implies that $v = \bar{x}$ and

$$d(\bar{x}, x^*) = \lim_{k\to\infty} d(x_{n_k}, x^*) \tag{31}$$

As in the proof of the previous proposition, convergence of $\{x_{n_k}\}_{k\in\mathbb{N}}$ to $\bar{x}$ with respect to the distance d, together with convergence in the weak topology yield strong convergence in L^1.

Consequently, the whole sequence $\{x_n\}_{n\in\mathbb{N}}$ converges strongly to $\bar{x}$ and thus the proof of the theorem is completed. □

5.3. *The symmetric KL-functional $J(u,v)$ as a regularizor*. In [7], the other directional entropy $d(x^*,\cdot)$ plays the role of the penalty term, motivated by the possibility of employing an EM algorithm in a discrete manner. The question of whether the symmetric entropic divergence $J(\cdot,x^*)=d(\cdot,x^*)+d(x^*,\cdot)$ is a more suitable regularizor arises naturally. Note that all the results above hold when $d(\cdot,x^*)$ is replaced by $J(\cdot,x^*)$. Thus, one can choose this entropic functional for regularization. Advantages associated with its use have been discussed in Section 2.

6. CONVERGENCE RATES

Convergence rates can be established under a suitable source condition involving the Banach space adjoint operator $A^*: L^\infty(\Sigma)\to L^\infty(\Omega)$.

Proposition
Let the following source condition hold:

$$A^*w=\ln\frac{\bar{x}}{x^*} \tag{32}$$

for some source element $w\in L^\infty(\Sigma)$. Then, for the choice $\alpha\sim\delta$, one has the convergence rate

$$\|x_\alpha^\delta-\bar{x}\|_1=O(\sqrt{\delta}) \tag{33}$$

Proof
Equality (32) implies that $\ln\bar{x}$ and $\bar{x}$ belong to $L^\infty(\Omega)$, because A^*w and $\ln x^*$ have the same property (see Remark in subsection 4.2). Then, by applying Lemma 3.3(iii), it follows that $\partial g(\bar{x})=\{1+\ln\bar{x}\}$ and that $d(x_\alpha^\delta,\bar{x})$ is finite. By (28), one obtains

$$\begin{aligned}\delta^2&\geqslant d(y^\delta,Ax_\alpha^\delta)+\alpha d(x_\alpha^\delta,x^*)-\alpha d(\bar{x},x^*)\\&=d(y^\delta,Ax_\alpha^\delta)+\alpha d(x_\alpha^\delta,\bar{x})+\alpha\langle A^*w,x_\alpha^\delta-\bar{x}\rangle\\&=d(y^\delta,Ax_\alpha^\delta)+\alpha d(x_\alpha^\delta,\bar{x})+\alpha\langle w,Ax_\alpha^\delta-A\bar{x}\rangle\end{aligned}$$

This implies that

$$d(y^\delta,Ax_\alpha^\delta)+\alpha d(x_\alpha^\delta,\bar{x})\leqslant\delta^2+\alpha\|w\|_\infty\|Ax_\alpha^\delta-A\bar{x}\|_1 \tag{34}$$

By inequality (13), there exists a positive constant a_1 such that $\|y^\delta-y\|_1^2\leqslant a_1 d(y^\delta,y)$, $\forall\delta>0$ sufficiently small. Consequently, the triangular norm inequality yields

$$\begin{aligned}\|Ax_\alpha^\delta-A\bar{x}\|_1^2&\leqslant 2(\|Ax_\alpha^\delta-y^\delta\|_1^2+\|y^\delta-y\|_1^2)\leqslant 2[\|Ax_\alpha^\delta-y^\delta\|_1^2+a_1 d(y^\delta,y)]\\&\leqslant 2(\|Ax_\alpha^\delta-y^\delta\|_1^2+a_1\delta^2)\end{aligned}$$

DOI: 10.1002/mma

By using the last inequality combined with (34) and (13), one obtains that

$$\begin{aligned}\tfrac{1}{2}\|Ax_\alpha^\delta - A\bar{x}\|_1^2 &\leqslant a_2 d(y^\delta, Ax_\alpha^\delta) + a_1\delta^2\\ &\leqslant (a_1+a_2)\delta^2 + \alpha a_2\|w\|_\infty \|Ax_\alpha^\delta - A\bar{x}\|_1 - \alpha a_2 d(x_\alpha^\delta, \bar{x})\end{aligned}$$

for some constant $a_2>0$ and for any $\delta>0$ sufficiently small. Then, one has

$$\tfrac{1}{2}\|Ax_\alpha^\delta - A\bar{x}\|_1^2 \leqslant (a_1+a_2)\delta^2 + \alpha a_2\|w\|_\infty \|Ax_\alpha^\delta - A\bar{x}\|_1 \tag{35}$$

and

$$\alpha d(x_\alpha^\delta, \bar{x}) \leqslant (a_1/a_2+1)\delta^2 + \alpha\|w\|_\infty \|Ax_\alpha^\delta - A\bar{x}\|_1 \tag{36}$$

The convergence rate $\|Ax_\alpha^\delta - A\bar{x}\|_1 = O(\delta)$ is immediately established from (35) for the choice $\alpha \sim \delta$. As a consequence of this and of inequality (36), it follows that $d(x_\alpha^\delta, \bar{x}) = O(\delta)$. In order to complete the proof, one applies again (13) and gets $\|x_\alpha^\delta - \bar{x}\|_1^2 \leqslant a_3 d(x_\alpha^\delta, \bar{x})$ for some $a_3>0$, which implies (33). □

ACKNOWLEDGEMENTS

The authors wish to thank Peter Hall (Australian National University) and Mark Westcott (CSIRO Mathematical and Information Sciences) for a number of valuable comments.

REFERENCES

1. Engl HW, Hanke M, Neubauer A. *Regularization of Inverse Problems*. Kluwer Academic Publishers: Dordrecht, 1996.
2. Tikhonov AN. Regularization of incorrectly posed problems. *Soviet Mathematics Doklady* 1963; **4**:1624–1627.
3. Twomey S. *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurement*. Elsevier: Amsterdam, 1977.
4. Wahba G. *Spline Models for Observational Data*. SIAM: Philadelphia, PA, 1990.
5. Latham GA, Anderssen RS. A hyperplane approach to the EMS algorithm. *Applied Mathematics Letters* 1992; **5**:71–74.
6. Vardi Y, Shepp LA, Kaufman L. A statistical model for positron emission tomography. *Journal of the American Statistical Association* 1985; **80**:8–37.
7. Iusem AN, Svaiter BF. A new smoothing-regularization approach for a maximum-likelihood estimation problem. *Applied Mathematics Optimization* 1994; **29**(3):225–241.
8. Besnerais GL, Bercher J-F, Demoment G. A new look at entropy for solving linear inverse problems. *IEEE Transactions on Information Theory* 1999; **45**:1566–1578.
9. Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**:79–86.
10. Hall P. Akaike's information criterion and Kullback–Leibler loss for histogram estimation. *Problems Theory and Related Fields* 1990; **85**:449–467.
11. Jeffreys H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, *Series A* 1946; **186**:453–461.
12. Halmos PR, Savage LJ. Application of the Radon–Nikodym theorem to the theory of sufficient statistics. *Annals of Mathematical Statistics* 1949; **20**:225–241.

DOI: 10.1002/mma

13. Shore JE, Johnson RW. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory* 1980; **26**:26–37.
14. Jaynes ET. Information theory and statistical mechanics. *Physical Review* 1957; **106**:620–630; *Physical Review* 1957; **108**:171–190.
15. Rubinstein RY, Kroese DP. *The Cross-Entropy Method*. Springer: Berlin, 2004.
16. Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, *Series B* 1966; **28**:131–142.
17. Bernado JM, Smith AFM. *Bayesian Theory*. Wiley: Chichester, 1994.
18. Braess D, Sauer T. Bernstein polynomials and learning theory. *Journal of Approximation Theory* 2004; **128**: 187–206.
19. Braess D, Forster J, Sauer T, Simon HU. How to achieve minimax expected Kullback–Leibler distance from an unknown finite distribution. *Lecture Notes in Artificial Intelligence*, vol. 2533. Springer: Berlin, 2002; 380–394.
20. Fisher RA. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *Series A* 1921; **222**:309–368.
21. Fisher RA. Theory of statistical estimation. *Proceedings of Cambridge Philosophical Society* 1925; **22**:700–725.
22. Kullback S. *Information Theory and Statistics*. Wiley: New York, 1959.
23. Bregman LM. The relaxation method for finding common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* 1967; **7**:200–217.
24. Csiszar I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics* 1991; **19**:2032–2066.
25. Jaynes ET. In *Papers on Probability*, *Statistics and Statistical Physics*, Rosenkrantz RD (eds). Reidel: Dordrecht, 1982.
26. Amato U, Hughes W. Maximum entropy regularization of Fredholm integral equations of the first kind. *Inverse Problems* 1991; **7**:793–803.
27. Engl HW, Landl G. Convergence rates for maximum entropy regularization. *SIAM Journal on Numerical Analysis* 1993; **30**:1509–1536.
28. Eggermont PPB. Maximum entropy regularization for Fredholm integral equations of the first kind. *SIAM Journal on Mathematical Analysis* 1993; **24**:1557–1576.
29. Donoho DL, Johnstone IM, Hoch JC, Stern AS. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society*, *Series B* 1992; **54**:41–81.
30. Landl GA, Anderssen RS. Non-negative differentially constrained entropy-like regularization. *Inverse Problems* 1996; **12**:35–53.
31. Bousquet O, Elisseff A. Stability and generalization. *Journal of Machine Learning Research* 2002; **2**:499–526.
32. Hall P. On Kullback–Leibler loss and density-estimation. *Annals of Statistics* 1987; **15**:1491–1519.
33. Hall P. On the use of compactly supported density estimates in problems of discrimination. *Journal of Multivariate Analysis* 1987; **23**:131–158.
34. Resmerita E. Regularization of ill-posed inverse problems in Banach spaces: convergence rates. *Inverse Problems* 2005; **21**:1303–1314.
35. Csiszar I, Tusnady G. Information geometry and alternative minimization procedures. *Statistics and Decisions Supplement* 1984; **1**:205–237.
36. Chen G, Teboule M. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming* 1994; **64**:81–101.
37. Burger M, Osher S. Convergence rates of convex variational regularization. *Inverse Problems* 2004; **20**:1411–1421.
38. OSullivan JA. Roughness penalties on finite domains. *IEEE Transactions on Image Processing* 1995; **4**:1258–1264.
39. Anderssen RS, Latham GA, Westcott M. Statistical methodology for inverse problems. *Mathematical and Computer Modelling* 1995; **22**:10–12.
40. Good IJ, Gaskins RA. Nonparametric roughness penalties for probability densities. *Biometriks* 1971; **58**:255–277.
41. Wand MP, Jones MC. *Kernel Smoothing*. Chapman & Hall: London, 1995.
42. Eggermont PPB, LaRiccia V. Maximum penalized likelihood estimation and smoothed EM algorithms for positive integral equations of the first kind. *Numerical Functional Analysis and Optimization* 1996; **17**:737–754.
43. Green PJ. On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society*, *Series B* 1990; **52**:443–452.
44. Dempster AP, Laird NM, Rubin DB. Maximum likelihood for incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.

DOI: 10.1002/mma

45. Yu S, Latham GA, Anderssen RS. Stabilizing properties of maximum penalized likelihood estimation for additive Poisson regression. *Inverse Problems* 1994; **10**:1199–1209.
46. Yu S, Latham GA, Anderssen RS. Matrix analysis of a two-stage-splitting iteration for maximum penalized likelihood estimation. *SIAM Journal on Matrix Analysis* 1997; **18**:348–359.
47. Lindblad G. Entropy, information and quantum measurements. *Communications in Mathematical Physics* 1973; **33**:305–322.
48. Borwein JM, Lewis AS. Convergence of best entropy estimates. *SIAM Journal on Optimization* 1991; **1**(2): 191–205.

DOI: 10.1002/mma