Majorization–Minimization Algorithms for Wavelet-Based Image Restoration

Mário A. T. Figueiredo, *Senior Member, IEEE*, José M. Bioucas-Dias, *Member, IEEE*, and Robert D. Nowak, *Senior Member, IEEE*

***Abstract*—Standard formulations of image/signal deconvolution under wavelet-based priors/regularizers lead to very high-dimensional optimization problems involving the following difficulties: the non-Gaussian (heavy-tailed) wavelet priors lead to objective functions which are nonquadratic, usually nondifferentiable, and sometimes even nonconvex; the presence of the convolution operator destroys the separability which underlies the simplicity of wavelet-based denoising. This paper presents a unified view of several recently proposed algorithms for handling this class of optimization problems, placing them in a common majorization–minimization (MM) framework. One of the classes of algorithms considered (when using quadratic bounds on nondifferentiable log-priors) shares the infamous "singularity issue" (SI) of "iteratively reweighted least squares" (IRLS) algorithms: the possibility of having to handle infinite weights, which may cause both numerical and convergence issues. In this paper, we prove several new results which strongly support the claim that the SI does not compromise the usefulness of this class of algorithms. Exploiting the unified MM perspective, we introduce a new algorithm, resulting from using ℓ_1 bounds for nonconvex regularizers; the experiments confirm the superior performance of this method, when compared to the one based on quadratic majorization. Finally, an experimental comparison of the several algorithms, reveals their relative merits for different standard types of scenarios.**

***Index Terms*—Image deconvolution, image restoration, majorization–minimization (MM) algorithms, optimization, regularization, wavelets.**

I. Introduction

WAVELET-BASED methods are the current state-of-the-art in image denoising, both in terms of performance and computational efficiency (see, e.g., [26], [42], [43], [45], [47], and the many references therein). However, image restoration in general (e.g., deblurring/deconvolution) is much more challenging than denoising, and applying wavelets turns out to be a much harder task. Unlike most approaches to wavelet-based denoising, which lead to thresholding rules, the optimization problems resulting from the wavelet-based formulations of deconvolution have no simple closed-form solutions (except in special circumstances [21]).

Most formulations of image deconvolution under wavelet-based priors lead to very large scale optimization problems where the objective function has two terms: a quadratic log-likelihood (or data discrepancy) term plus a (usually non quadratic) log-prior (also known as regularizer of penalty function). In addition to being of very large dimensionality, these optimization problems are also difficult for two other main reasons: the best performing penalty functions are nondifferentiable and sometimes even nonconvex; the presence of a convolution operator (rather than simply additive white Gaussian noise) destroys the separability which underlies the simplicity of wavelet-based denoising. These optimization problems have been recently addressed via expectation-maximization (EM) algorithms [7], [27], [28], as well as by majorization–minimization (MM) methods (also known as bound optimization or surrogate optimization methods; see [36] for a tutorial/review on MM algorithms) [18], [29]. Earlier approaches to wavelet-based image restoration were recently reviewed in [7] and [28], so we refrain from reviewing them here, and simply indicate some key references: [5], [6], [21], [40], [44].

This paper focuses on the class of MM approaches to wavelet-based image restoration by considering three possible majorization strategies leading to three different classes of algorithms, as described in the following three sections.

A. MM Algorithms via Majorizing the Log-Likelihood

We show that the methods independently introduced by several authors [18], [23], [24], [27], [28], [40], [49], [50] can all be seen as MM algorithms based on a separable quadratic majorizer on the log-likelihood. This class of algorithms involve the iterative application of nonlinear shrinkage/thresholding denoising operators; thus, they are termed *iterative shrinkage-thresholding* (IST) or *iterative denoising* algorithms. Convergence proofs for this class of algorithms have been recently presented in [16] and [18].

B. MM Algorithms via Majorizing the Penalty Function

When a quadratic separable majorizer on the penalty function is adopted, the resulting MM algorithm has the structure of an *iteratively reweighted shrinkage* (IRS) which is related to the well known reweighted least squares (IRLS) algorithm. In the context of wavelet-based image restoration, this scheme was introduced in [7] using an EM framework.

Algorithms of the IRLS type have been often criticized due to what can be called the "singularity issue": when using quadratic majorizers for nondifferentiable functions, if, at some iteration,

Manuscript received January 11, 2007; revised August 11, 2007. This work was supported in part by Fundação para a Ciência e Tecnologia (FCT), Portuguese Ministry of Science and Higher Education, under project POSC/EEA-CPS/61271/2004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Eric Kolaczyk.

M. A. T. Figueiredo and J. M. Bioucas-Dias are with the Instituto de Telecomunicações and the Instituto Superior Técnico, Technical University of Lisbon, 1049-001 Lisboa, Portugal (e-mail: mario.figueiredo@lx.it.pt; jose.bioucas@lx.it.pt).

R. D. Nowak is with the Electrical and Computer Engineering Department, University of Wisconsin, Madison, WI 53706 USA (e-mail: nowak@ece.wisc.edu).

Digital Object Identifier 10.1109/TIP.2007.909318

one of the variables coincides with a point of nondifferentiability, the corresponding weight is infinity, thus locking this variable at that point. This effect raises numerical difficulties (handling infinity) and may prevent convergence of the algorithm.

In this paper, we show several new results concerning the infamous "singularity issue," which strongly suggest that this issue does not compromise the usefulness of this class of algorithms. More specifically, we show the following.

a) The algorithm can be written in such a way that it dispenses with having to handle infinite values.
b) If initialized with all components different from zero, then, with probability one, no component will become zero in a finite number of iterations.
c) If the algorithm converges, it does so to a minimizer of the objective function (with probability one).

C. MM Algorithms via Majorizing Both the Log-Likelihood and the Penalty Function

We introduce a new class of MM algorithms, obtained by combining the separable quadratic majorizer on the log-likelihood with a majorizer on penalty function, for which we consider two options: with a quadratic majorizer, we recover a particular instance of the algorithm introduced in [7]; with an ℓ_1 majorizer, which is well suited for nonconvex penalty functions, we obtain a new class of algorithms which we call *iterative soft* thresholding (ISoft).

D. Outline of the Paper

The remaining sections of the paper are organized as follows. Section II reviews the formulation of wavelet-based image restoration as an optimization problem, analyzes the sources of the difficulties in handling that optimization problem, and mentions related work. Section III contains a brief introduction to MM algorithms. In Section IV, a class of MM algorithm is derived by considering majorizers on the log-likelihood term of the objective function. Another class of algorithms, obtained by using majorizers on the penalty function, is presented in Section V; that section also contain new theoretical results concerning the properties of this class of algorithms. In Section VII, we summarize the algorithms and analyze their computational cost per iteration. Section VIII presents an experimental comparison of the several types of algorithms, showing their relative merits for different types of scenarios, in terms of: severity of the blur operator; amount of added noise; nature of the adopted prior. Finally, Section IX ends the paper with some concluding remarks.

II. Problem Formulation

A. Wavelet-Based Image Deconvolution

In this paper, we adopt the standard convention of representing images as vectors, obtained by stacking all the pixels in some predetermined order (e.g., lexicographically). In image reconstruction/restoration problems, the goal is to estimate an original image $\mathbf{x}$ from an observation $\mathbf{y}$, assumed to have been produced by the linear-Gaussian observation model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{1}$$

where matrix $\mathbf{H}$ represents the observation operator, and $\mathbf{n}$ contains samples of independent zero-mean Gaussian random variables of variance σ^2. Matrix $\mathbf{H}$ can model many types of linear observations, but this paper will focus on deconvolution (e.g., deblurring) problems. In this case, matrix $\mathbf{H}$ represents a 2-D convolution and it is block-circulant with circulant blocks (assuming periodic boundary conditions for the convolution) or block Toeplitz with Toeplitz blocks [1]. Multiplying any vector (image) by $\mathbf{H}$ or $\mathbf{H}^T$ can, thus, be done using the 2-D fast Fourier transform (FFT) with a cost of $O(N \log N)$, where N is the number of image pixels.

To obtain a wavelet-based formulation, consider that $\mathbf{x}$ can be represented on some wavelet basis as $\mathbf{x} = \mathbf{W}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the vector of representation coefficients and the set of columns of $\mathbf{W}$ is a wavelet basis or dictionary. In the case of an orthogonal basis, $\mathbf{W}$ is a square orthogonal matrix, whereas for an over-complete dictionary (e.g., a tight frame), $\mathbf{W}$ has more columns than rows. With this wavelet-based representation, the observation model becomes

$$\mathbf{y} = \mathbf{H}\mathbf{W}\boldsymbol{\theta} + \mathbf{n} \tag{2}$$

and the resulting log-likelihood function is

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{H}\mathbf{W}\boldsymbol{\theta}\|_2^2 + K \tag{3}$$

where $\|\cdot\|_2^2$ denotes the usual squared Euclidean norm and K is a constant independent of $\boldsymbol{\theta}$.

The *maximum penalized likelihood* (MPL) estimate of $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{4}$$

where

$$L(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{W}\boldsymbol{\theta}\|_2^2 + \lambda C(\boldsymbol{\theta}) \tag{5}$$

where $C(\boldsymbol{\theta})$ is a penalty function which has several different possible interpretations, depending on the framework in which the problem is formulated. In Bayesian decision theoretic terms, (4) defines the well-known maximum *a posteriori* (MAP) estimate, with $\lambda C(\boldsymbol{\theta}) = -\sigma^2 \log p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is a prior density (usually heavy-tailed), expressing the sparse nature of the wavelet coefficients of natural images [43]. The estimation criterion (4) can also be seen in a regularization perspective as a way to address the ill-posed problem of inferring $\boldsymbol{\theta}$ from $\mathbf{y}$; in that setting, $C(\boldsymbol{\theta})$ is called the regularization function and λ is the regularization parameter [3].

Of course, the MAP/MPL criterion is not the only possible choice for wavelet-based image denoising/restoration, and several alternatives have been proposed with excellent results [33], [44], [45], [48]. In this paper, we are solely concerned with algorithms for solving the MAP/MPL criterion, and will not discuss the relative merits of this option with respect to the possible alternatives.

B. Gaussian Priors/Quadratic Penalties

The simplest version of (4) is obtained when a zero-mean Gaussian prior for $\boldsymbol{\theta}$ is adopted

$$-\sigma^2 \log p(\boldsymbol{\theta}) = \lambda C(\boldsymbol{\theta}) = \frac{\lambda}{2}\boldsymbol{\theta}^T\mathbf{P}\boldsymbol{\theta} + R$$

where $\mathbf{P}$ is symmetric positive semi-definite and R is a scalar independent of $\boldsymbol{\theta}$. In this case, the solution of (4) is

$$\widehat{\boldsymbol{\theta}} = (\mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W} + \lambda\mathbf{P})^{-1}\mathbf{W}^T\mathbf{H}^T\mathbf{y}. \tag{6}$$

Of course, this estimate can only be obtained via an iterative algorithm, due to the huge size of the matrix being inverted; in fact, it is not even practical to explicitly compute it or store it (e.g., for 256×256 images, it would be a $256^2 \times 256^2$ matrix).

C. Non-Gaussian (Sparseness-Inducing) Priors

It is well accepted that Gaussian densities are not adequate models for the statistics of wavelet coefficients of natural images; the sparse nature of wavelet-based representations (many very small or even zero coefficients together with a few very large ones) demands heavy-tailed densities [41]. One of the distributions most often adopted to model the statistics of wavelet coefficients is the independent generalized Gaussian density (GGD, see [43])

$$p(\boldsymbol{\theta}) \propto \exp\left\{-\tau \sum_i |\theta_i|^p\right\}. \tag{7}$$

The logarithm of this prior is proportional to the pth power of an ℓ_p norm[1] plus an irrelevant constant S, that is

$$-\sigma^2 \log p(\boldsymbol{\theta}) = \lambda C(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_p^p + S$$

where $\lambda = \sigma^2\tau$. It has been found that good wavelet-based models of natural images are obtained for $p < 1$ [43].

Another class of heavy-tailed prior densities which has been used to model wavelet coefficients (and which contains GGDs with $p \leq 2$ as special cases) is that of Gaussian scale mixtures (GSM); see [2], [7], [17], and [45] for details.

If (4) is hard to solve when $p(\boldsymbol{\theta})$ is a Gaussian prior, it becomes much harder when $p(\boldsymbol{\theta})$ is a heavy-tailed prior, such as a GGD or GSM. In this case, we no longer even have a "closed-form" expression [such as (6)].

D. Sources of Difficulties

The difficulty of solving (4) has two main sources.

- Matrix $\mathbf{HW}$, unlike $\mathbf{H}$ alone, is not block-circulant (nor block-Toeplitz), thus, cannot be efficiently handled using FFT-based methods. Even when $\mathbf{W}$ is orthogonal, $\mathbf{HW}$ is not. The presence of this matrix makes solving (4), even in the Gaussian case examined in Section II-B, a task that can only be achieved using iterative algorithms.
- When the penalty $C(\boldsymbol{\theta})$ (equivalently, the log prior $-\log p(\boldsymbol{\theta})$) is not a quadratic function of $\boldsymbol{\theta}$, there is, in general,[2] no close-form solution to (4).

In this paper, we will describe MM algorithms which are obtained by addressing each one (or both) of this difficulties; that is, by using majorizers for the log-likelihood or/and the penalty function.

[1]Recall that the ℓ_p norm is defined as $\|\mathbf{v}\|_p = (\sum_i |v_i|^p)^{1/p}$; thus $\|\mathbf{v}\|_p^p = \sum_i |v_i|^p$. Although, for $p < 1$, $\|\mathbf{v}\|_p$ is not a norm, we will (as is commonly done) still refer to it as a norm.

[2]Of course, if $\mathbf{H} = \mathbf{I}$ and $\mathbf{W}$ is orthogonal, (4) may have closed-form solution for some choices of $p(\boldsymbol{\theta})$; however, in this case, we would be in the presence of a pure denoising problem, not a deconvolution problem.

E. Related Problems and Approaches

Optimization problems formally close to (4), with $\mathbf{HW}$ replaced by some arbitrary matrix $\mathbf{A}$, have been studied in other contexts and applications. For example, with $\mathbf{A}$ being the *design matrix* of some regression problem, the LASSO (least absolute shrinkage and selection operator) criterion is similar to (4), with $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ [51]. Notice, however, that state-of-the-art algorithms which have been proposed to solve the LASSO (such as *least angle regression* [22]) cannot be used to address (4) because matrix $\mathbf{HW}$ can not be explicitly computed or stored, nor is it possible to access individual rows, columns, or elements. This fact places (4) beyond the reach of most general-purpose optimization methods.

Another problem formulation leading to an objective function with the same form as (4) is the following. Let the columns of $\mathbf{W}$ contain a redundant (over-complete) dictionary with respect to which a representation of the observed image (or signal) is sought [14], [23], [24]. This representation can be obtained by solving (4), with $\mathbf{H} = \mathbf{I}$ and $C(\boldsymbol{\theta})$ being some penalty function encouraging sparse solutions [23], [24]. The algorithms considered in this paper can be directly applied to this scenario. For $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, this is known as the *basis-pursuit denoising* problem [14].

Finally, we should mention that MM algorithms have been used for more than a decade in image reconstruction (mainly in tomographic medical imaging, see, e.g., [20], [25], and [39]). However, to the best of our knowledge, they have only very recently been used to tackle the optimization problems that result from wavelet-based approaches to inverse problems (e.g., deconvolution) [18], [28], [29].

III. Majorization–Minimization Algorithms

A MM [36] iterative algorithm for solving (4) has the form

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)}\right) \tag{8}$$

where $Q(\boldsymbol{\theta}; \boldsymbol{\theta}') \geq L(\boldsymbol{\theta})$, for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$, and $Q(\boldsymbol{\theta}; \boldsymbol{\theta}) = L(\boldsymbol{\theta})$, i.e., $Q(\boldsymbol{\theta}; \boldsymbol{\theta}')$ upper bounds (majorizes) $L(\boldsymbol{\theta})$, touching it for $\boldsymbol{\theta} = \boldsymbol{\theta}'$. It is well known that this property of the Qfunction implies monotonicity of the algorithm, since

$$\begin{aligned} L\left(\widehat{\boldsymbol{\theta}}^{(t+1)}\right) &= L\left(\widehat{\boldsymbol{\theta}}^{(t+1)}\right) - Q\left(\widehat{\boldsymbol{\theta}}^{(t+1)}; \widehat{\boldsymbol{\theta}}^{(t)}\right) \\ &\quad + Q\left(\widehat{\boldsymbol{\theta}}^{(t+1)}; \widehat{\boldsymbol{\theta}}^{(t)}\right) \\ &\leq Q\left(\widehat{\boldsymbol{\theta}}^{(t+1)}; \widehat{\boldsymbol{\theta}}^{(t)}\right) \\ &\leq Q\left(\widehat{\boldsymbol{\theta}}^{(t)}; \widehat{\boldsymbol{\theta}}^{(t)}\right) = L\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) \end{aligned} \tag{9}$$

where the first inequality results from $Q(\boldsymbol{\theta}; \boldsymbol{\theta}') \geq L(\boldsymbol{\theta})$, the second one from the fact that, according to (8), $Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$ attains its minimum for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t+1)}$.

The MM approach opens the door to the derivation of EM-type algorithms [19], where the Q-function (the majorizer) does not have to result from a model with missing-data, as in

standard EM. Any convenient inequality and any property of $L(\boldsymbol{\theta})$ can be invoked to obtain a valid Q-function [36].

MM algorithms have three properties (which have trivial proofs), of which we will make use later.

- **Property 1:** The function $Q_a(\boldsymbol{\theta};\boldsymbol{\theta}') = AQ(\boldsymbol{\theta};\boldsymbol{\theta}') + B$, where $A > 0$ and B are constants independent of $\boldsymbol{\theta}$ (possibly dependent on $\boldsymbol{\theta}'$) defines exactly the same iteration as $Q(\boldsymbol{\theta};\boldsymbol{\theta}')$.
- **Property 2:** Let $L(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}) + L_2(\boldsymbol{\theta})$; consider two majorizers, $Q_1(\boldsymbol{\theta};\boldsymbol{\theta}') \geq L_1(\boldsymbol{\theta})$ and $Q_2(\boldsymbol{\theta};\boldsymbol{\theta}') \geq L_2(\boldsymbol{\theta})$, both with equality for $\boldsymbol{\theta} = \boldsymbol{\theta}'$. Then, all the following functions majorize $L(\boldsymbol{\theta})$ (with equality for $\boldsymbol{\theta} = \boldsymbol{\theta}'$): $Q_1(\boldsymbol{\theta};\boldsymbol{\theta}') + Q_2(\boldsymbol{\theta};\boldsymbol{\theta}')$, $L_1(\boldsymbol{\theta}) + Q_2(\boldsymbol{\theta};\boldsymbol{\theta}')$, and $Q_1(\boldsymbol{\theta};\boldsymbol{\theta}') + L_2(\boldsymbol{\theta})$.
- **Property 3:** The monotonicity property of MM is kept if, instead of exactly minimizing $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(t)})$ [as in (8)], the following weaker condition is satisfied:

$$\widehat{\boldsymbol{\theta}}^{(t+1)} \text{ is such that } Q\left(\widehat{\boldsymbol{\theta}}^{(t+1)};\widehat{\boldsymbol{\theta}}^{(t)}\right) \leq Q\left(\widehat{\boldsymbol{\theta}}^{(t)};\widehat{\boldsymbol{\theta}}^{(t)}\right). \quad (10)$$

Notice that this is the only property of $\widehat{\boldsymbol{\theta}}^{(t+1)}$ that was invoked in showing the monotonicity of MM. A similar reasoning underlies *generalized* EM (GEM) algorithms [53]. Algorithms defined by iteration (10), instead of (8) are, thus, called *generalized* MM (GMM) algorithms.

IV. MM Algorithm via Majorization of the Log-Likelihood

A. Majorizing the Log-Likelihood

Let us denote $L_\ell(\boldsymbol{\theta}) = (1/2)\|\mathbf{y} - \mathbf{H}\mathbf{W}\boldsymbol{\theta}\|^2$, the log-likelihood term of the objective function in (5). This is a quadratic function with positive semi-definite Hessian $\mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W}$, thus convex (though not necessarily strictly so), and gradient $\mathbf{W}^T\mathbf{H}^T(\mathbf{H}\mathbf{W}\boldsymbol{\theta} - \mathbf{y})$. We can write a second-order Taylor expansion of this function (which is exact, because the function is quadratic) about some point $\boldsymbol{\theta}'$

$$L_\ell(\boldsymbol{\theta}) = L_\ell(\boldsymbol{\theta}') + (\boldsymbol{\theta} - \boldsymbol{\theta}')^T \underbrace{\mathbf{W}^T\mathbf{H}^T(\mathbf{H}\mathbf{W}\boldsymbol{\theta}' - \mathbf{y})}_{\text{gradient at } \boldsymbol{\theta}'} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \underbrace{\mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W}}_{\text{Hessian}}(\boldsymbol{\theta} - \boldsymbol{\theta}'). \quad (11)$$

Now let $\mathbf{D}$ be a symmetric matrix such that

$$\mathbf{D} \succeq \mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W} \quad (12)$$

where $\succeq$ denotes matrix inequality.[3] Since $\mathbf{A} \succeq \mathbf{B} \Rightarrow \mathbf{v}^T(\mathbf{A} - \mathbf{B})\mathbf{v} \geq 0$, for any $\mathbf{v}$, we can obtain a majorizer for $L_\ell(\boldsymbol{\theta})$ as

$$L_\ell(\boldsymbol{\theta}) \leq L_\ell(\boldsymbol{\theta}') + (\boldsymbol{\theta} - \boldsymbol{\theta}')^T\mathbf{W}^T\mathbf{H}^T(\mathbf{H}\mathbf{W}\boldsymbol{\theta}' - \mathbf{y}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T\mathbf{D}(\boldsymbol{\theta} - \boldsymbol{\theta}') \quad (13)$$

which is, of course, an equality for $\boldsymbol{\theta} = \boldsymbol{\theta}'$. This suggests using the r.h.s. of (13) as $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(t)})$, with $\boldsymbol{\theta}' = \widehat{\boldsymbol{\theta}}^{(t)}$. According to [36], this quadratic bounding approach to obtaining a monotonic algorithm was first introduced in [8].

A choice of $\mathbf{D}$ leading to a simple algorithm is a matrix proportional to identity. In fact, as stated by the following proposition (shown in Appendix A1) $\mathbf{I} \succeq \mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W}$, meaning that we can use $\mathbf{D} = \mathbf{I}$ in (13).

Proposition 1: Let the set of columns of $\mathbf{W}$ correspond to a normalized tight frame, that is,[4] $\mathbf{W}\mathbf{W}^T = \mathbf{I}$ and $\mathbf{H}$ be normalized such that $\|\mathbf{H}\|_2 = 1$. Then, $\mathbf{I} \succeq \mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W}$.

Inserting $\mathbf{D} = \mathbf{I}$ into (13), we can write (after some simple manipulation)

$$L_\ell(\boldsymbol{\theta}) \leq \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\phi}^{(t)}\right\|_2^2 + K \quad (14)$$

where K is a constant independent of $\boldsymbol{\theta}$ and

$$\boldsymbol{\phi}^{(t)} = \widehat{\boldsymbol{\theta}}^{(t)} + \mathbf{W}^T\mathbf{H}^T\left(\mathbf{y} - \mathbf{H}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)}\right). \quad (15)$$

B. Update Rule

With a majorizer for L_ℓ in hand, we invoke **Property 1** to drop K and **Property 2** to use (14) to build a majorizer for the complete objective function $L_\ell(\boldsymbol{\theta}) + \lambda C(\boldsymbol{\theta})$. The resulting update equation is, thus

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\min_{\boldsymbol{\theta}}\left\{\frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\phi}^{(t)}\right\|_2^2 + \lambda C(\boldsymbol{\theta})\right\}. \quad (16)$$

Notice that (16) corresponds to a pure denoising problem [the same as (4) and (5), with $\mathbf{H}\mathbf{W} = \mathbf{I}$], under a penalty/log-prior $C(\boldsymbol{\theta})$, and with "noisy coefficients" $\boldsymbol{\phi}^{(t)}$. Denoting as $\boldsymbol{\Psi}_{C,\lambda}$ the function which returns the solution of (16), which is a so-called "denoising rule," we can write (16) as

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \boldsymbol{\Psi}_{C,\lambda}\left(\boldsymbol{\phi}^{(t)}\right) = \boldsymbol{\Psi}_{C,\lambda}\left(\widehat{\boldsymbol{\theta}}^{(t)} + \mathbf{W}^T\mathbf{H}^T\left(\mathbf{y} - \mathbf{H}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)}\right)\right). \quad (17)$$

The algorithm defined by (17), termed *iterative shrinkage-thresholding* (IST), coincides with those previously presented in [18], [28], and [29]. Theoretical results concerning the convergence of this iterative procedure can be found in [18], for the case of convex GGD priors, that is, for $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_p^p$, with $1 \leq p \leq 2$. The results in [18] were recently extended and generalized in [16]. Similar algorithms were also proposed in [49] and [50], without any formal support or analysis, but with excellent practical results. Algorithms of the same class were also proposed in [23] and [24], to find sparse representations on redundant dictionaries.

For a few choices of $C(\boldsymbol{\theta})$, there are closed-form expressions for $\boldsymbol{\Psi}_{C,\lambda}$. We focus only on decoupled penalty functions of the form $C(\boldsymbol{\theta}) = \sum_i C(\theta_i)$. In this case, (16) can be solved separately w.r.t. each component

$$\widehat{\theta}_i^{(t+1)} = \arg\min_{\theta_i}\left\{\frac{1}{2}(\theta_i - \phi_i)^2 + \lambda C(\theta_i)\right\} \quad (18)$$

[3]Recall that $\mathbf{A} \succeq \mathbf{B}$ (for two symmetric matrices) means that matrix $\mathbf{A} - \mathbf{B}$ is positive semi-definite.

[4]If the columns of a matrix correspond to a normalized tight frame, then $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, but $\mathbf{W}^T\mathbf{W}$ may be different from identity, because $\mathbf{W}$ may not be orthogonal; see [10] and [41] for an introduction to frames.

where ϕ_i denotes the ith component of $\boldsymbol{\phi}^{(t)}$. There are two standard cases for which (18) has simple closed-form solutions. For a zero-mean Gaussian prior, $C(\theta_i) = (1/2)\theta_i^2$, the solution is simply

$$\widehat{\theta}_i^{(t+1)} = \frac{\phi_i}{1+\lambda}. \tag{19}$$

For a Laplacian prior (i.e., $C(\theta_i) = |\theta_i|^p$ with $p = 1$), we have

$$\widehat{\theta}_i^{(t+1)} = \text{soft}(\phi_i, \lambda) = \text{sign}(\phi_i)\max\{0, |\phi_i| - \lambda\}$$

the well-known *soft threshold* (ST) function [43]. The closed-form solution of (18), with $C(\theta_i) = |\theta_i|^p$, also exists for $p \in \{4/3, 3/2, 3, 4\}$ [13]. Finally, the also popular hard-threshold (HT) function can be seen as the limit of (18), with $C(\theta_i) = |\theta_i|^p$, when p goes to zero (see [43] for details).

A shrinkage/thresholding function which was shown in [7] and [28] to be very effective for wavelet-based deconvolution is the *non-negative garrote* (NNG)

$$\widehat{\theta}_i^{(t+1)} = \text{garrote}(\phi_i, \lambda) = \frac{\max\{0, \phi_i^2 - \lambda^2\}}{\phi_i + 1_{\phi_i=0}} \tag{20}$$

where $1_{a=0}$ is the indicator function of the condition $a = 0$. As shown in [7], the NNG corresponds to the solution of (18) under a prior which does belong to the GSM family.

V. MM Algorithm via Majorization of the Penalty

A. Majorizing the Penalty/Log-Prior

In this section, we derive MM algorithms by considering majorizers for GGD (for $0 < p \leq 2$) and GSM log-priors. We consider only independent priors, where $p(\boldsymbol{\theta}) = \prod_i p(\theta_i)$ (equivalently, $C(\boldsymbol{\theta}) = \sum_i C(\theta_i)$), where the marginal densities $p(\theta_i)$ belong to a GSM family (of which GGDs are a particular case). Even in denoising problems (where $\mathbf{H} = \mathbf{I}$) with an orthogonal wavelet basis ($\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$), which allows decoupling the solution of (4), most priors in this class do not lead to closed-form solutions (except in a few cases mentioned in Section IV-B).

Let us take note of some properties of GSMs which will be needed below. Any (univariate) GSM density $p(\theta)$ is necessarily even, since it is a convex combination (maybe infinite) of even functions (zero-mean Gaussian densities). For the same reason, any (univariate) GSM density $p(\theta)$ is a decreasing function of $|\theta|$, thus $\lambda C(\theta) = -\sigma^2 \log p(\theta)$ is an increasing function of $|\theta|$, of course also even. Since GSMs have heavier tails than a pure Gaussian, the corresponding penalty $\lambda C(\theta) = -\sigma^2 \log p(\theta)$ necessarily grows slower than a quadratic function. Finally, since $p(\theta)$ is a GSM, both $p(\theta)$ and $C(\theta)$ are C^∞, except maybe at the origin [32], [46].

Since $C(\theta)$ is even and subquadratic, it is majorized by an even quadratic function; i.e., we seek η and ν such that

$$\lambda C(\theta) \leq \frac{\eta}{2}\theta^2 + \nu \tag{21}$$

with equality for $\theta = \theta'$, where $\theta' = \widehat{\theta}^{(t)}$ denotes the previous iterate, all throughout this section. This requires $(\eta/2)\theta^2 + \nu$ to be tangent to $\lambda C(\theta)$ at θ', that is, their derivatives at θ' must coincide. This condition leads to

$$\eta = \lambda \frac{C'(\theta')}{\theta'} \equiv \Upsilon(\theta') \tag{22}$$

where $C'(\theta)$ is the derivative of $C(\cdot)$ at θ. Of course, we could also solve for ν to have the majorizer touch $\lambda C(\theta)$ at θ', but this value is irrelevant for the algorithm (see **Property 1**).

Notice that when the penalty corresponds to a log-prior, $\lambda C(\theta) = -\sigma^2 \log p(\theta)$, (22) can be written as

$$\Upsilon(\theta') = -\sigma^2 \frac{1}{\theta'} \left.\frac{d \log p(\theta)}{d\theta}\right|_{\theta'} = -\sigma^2 \frac{p'(\theta')}{\theta' p(\theta')}$$

which coincides with equation (18) in [7]; this shows the method therein derived under an EM framework, also has an MM interpretation, based on quadratic majorizers for GSM log-priors. This quadratic bounding technique is well known in robust regression, where it is used to derive the iteratively reweighted least squares (IRLS) method [35].

Notice that $\Upsilon(\theta')$ in (22) is not defined for $\theta' = 0$. If $C(\theta)$ has finite second derivative at the origin, we can define $\Upsilon(0)$ by continuity. Noticing that, in this case, $C'(0) = 0$, we have

$$\lim_{\theta \to 0} \frac{C'(\theta)}{\theta} = \lim_{\theta \to 0} \frac{C'(\theta) - C'(0)}{\theta - 0} = C''(0)$$

by definition of second derivative, which is by hypothesis finite. In this case, the objective function is strictly convex and twice differentiable, and the Q-function

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}') = \|\mathbf{y} - \mathbf{H}\mathbf{W}\theta\|^2 + \boldsymbol{\theta}^T \mathbf{D}(\boldsymbol{\theta}')\boldsymbol{\theta} \tag{23}$$

where $\mathbf{D}(\boldsymbol{\theta}') = \text{diag}(\Upsilon(\theta_i'), i = 1, 2, \ldots)$, is smooth. Thus, convergence of the resulting MM algorithm can be easily shown, following the same line of reasoning used to show convergence of EM [53].

However, the most often used penalties in wavelet-based image restoration are nondifferentiable at the origin, which is a sufficient condition for leading to sparse estimates [43]. For these penalties, we have to follow a different route. The function $q(\cdot;\cdot) : \mathbb{R}^2 \to \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, given by

$$q(\theta; \theta') = \begin{cases} \theta^2 \Upsilon(\theta')/2 + \nu, & \Leftarrow \quad \theta' \neq 0 \\ +\infty, & \Leftarrow \quad \theta' = 0 \wedge \theta \neq 0 \\ 0, & \Leftarrow \quad \theta' = 0 \wedge \theta = 0 \end{cases} \tag{24}$$

is well defined for all θ and θ', and is a valid majorizer because it satisfies $q(\theta, \theta') \geq \lambda C(\theta)$, with equality for $\theta = \theta'$. Finally, since $C(\boldsymbol{\theta}) = \sum_i C(\theta_i)$, we invoke **Property 2** to add the individual majorizers yielding the majorizer

$$\sum_i q(\theta_i; \theta_i') \geq \lambda C(\boldsymbol{\theta}) \tag{25}$$

with equality for $\boldsymbol{\theta} = \boldsymbol{\theta}'$. Adding this majorizer to the log-likelihood term $L_\ell(\boldsymbol{\theta}) = (1/2)\|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2$, where $\mathbf{A} = \mathbf{H}\mathbf{W}$, yields the Q-function

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}') = \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2 + \sum_i q(\theta_i; \theta_i'). \tag{26}$$

B. Update Rule

The updated iterate $\widehat{\boldsymbol{\theta}}^{(t+1)}$, denoted simply as $\boldsymbol{\theta}''$, is the minimizer of $Q(\boldsymbol{\theta};\boldsymbol{\theta}')$. The bound defined in (24) implies that the updating rule satisfies

$$(\theta_i'' = 0) \Leftarrow (\theta_i' = 0) \tag{27}$$

meaning that it can be stated as the constrained problem

$$\boldsymbol{\theta}'' = \arg\min_{\boldsymbol{\theta}} \left\{ \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2 + 2\sum_i q(\theta_i;\theta_i') \right\} \quad \text{subject to } \boldsymbol{\theta}_Z = \mathbf{0} \tag{28}$$

where $Z = \{i : \theta_i' = 0\}$, and $\boldsymbol{\theta}_Z$ is the subvector of $\boldsymbol{\theta}$ corresponding to the indices in Z. Letting $\widetilde{Z} = \{i : \theta_i' \neq 0\}$, we denote as $\mathbf{A}_{\widetilde{Z}}$ the matrix formed by the columns of $\mathbf{A}$ with indices in $\widetilde{Z}$. Problem (28) is equivalent to

$$\boldsymbol{\theta}_Z'' = \mathbf{0} \tag{29}$$

$$\boldsymbol{\theta}_{\widetilde{Z}}'' = \arg\min_{\boldsymbol{\theta}} \left\{ \|\mathbf{y} - \mathbf{A}_{\widetilde{Z}}\boldsymbol{\theta}\|^2 + \boldsymbol{\theta}^T \mathbf{D}_{\widetilde{Z}} \boldsymbol{\theta} \right\} \tag{30}$$

where $\mathbf{D}_{\widetilde{Z}} = \operatorname{diag}(\Upsilon(\theta_i'), i \in \widetilde{Z})$. Since (30) is quadratic, the minimizer is simply given by

$$\boldsymbol{\theta}_{\widetilde{Z}}'' = \left(\mathbf{A}_{\widetilde{Z}}^T \mathbf{A}_{\widetilde{Z}} + \mathbf{D}_{\widetilde{Z}} \right)^{-1} \mathbf{A}_{\widetilde{Z}}^T \mathbf{y}. \tag{31}$$

As shown Appendix A1, the update rule which combines (29) and (31) can be written compactly as

$$\boldsymbol{\theta}'' = \mathbf{E}\mathbf{A}^T (\mathbf{A}\mathbf{E}\mathbf{A}^T + \mathbf{I})^{-1} \mathbf{y} \tag{32}$$

where $\mathbf{E}$ is a diagonal matrix with the $E_{i,i}$ entry given by

$$E_{i,i} = \begin{cases} (\Upsilon(\theta_i'))^{-1}, & \Leftarrow \quad \theta_i' \neq 0 \\ 0, & \Leftarrow \quad \theta_i' = 0. \end{cases} \tag{33}$$

This form of the update equation shows that it is never necessary to handle infinite values, which is usually pointed out as a weakness of IRLS type algorithms. If a component becomes zero, the corresponding element of $\mathbf{E}$ also simply becomes zero. Of course, this will lock this component at zero forever, which may impact the convergence of the algorithm to a minimizer of the objective function. This issue will be analyzed in detail as follows in Section V-D.

C. Solving the Update Equation

To implement each update step, one can simply keep at zero the components that were zero and compute the remaining ones by solving (31). Of course, this does not require inverting the matrix, but just solving the corresponding system $(\mathbf{A}_{\widetilde{Z}}^T \mathbf{A}_{\widetilde{Z}} + \mathbf{D}_{\widetilde{Z}})\boldsymbol{\theta} = \mathbf{A}_{\widetilde{Z}}^T \mathbf{y}$. Due to its size, this system can only be solved iteratively. The approach proposed in [7] consists in using a *second-order* (also known as *two-step*) *stationary iterative method* (SOSIM) [4], which is defined by

$$\boldsymbol{\theta}_{\widetilde{Z}}^{(i+1)} = (\alpha - \beta)\boldsymbol{\theta}_{\widetilde{Z}}^{(i)} + (1-\alpha)\boldsymbol{\theta}_{\widetilde{Z}}^{(i-1)} + \beta[\mathbf{D}_{\widetilde{Z}} + \mathbf{I}]^{-1} \left[\boldsymbol{\theta}_{\widetilde{Z}}^{(i)} + \mathbf{A}_{\widetilde{Z}}^T \left(\mathbf{y} - \mathbf{A}_{\widetilde{Z}} \boldsymbol{\theta}_{\widetilde{Z}}^{(i)} \right) \right]. \tag{34}$$

Notice that the iteration counter i in (34) defines an inner loop (the SOSIM scheme) which is nested inside the MM iteration. Finally, the pair of update equations $\boldsymbol{\theta}_Z^{(i+1)} = \mathbf{0}$ and (34) can be written compactly as

$$\boldsymbol{\theta}^{(i+1)} = (\alpha - \beta)\boldsymbol{\theta}^{(i)} + (1-\alpha)\boldsymbol{\theta}^{(i-1)} + \beta \mathbf{E}[\mathbf{E} + \mathbf{I}]^{-1} \left[\boldsymbol{\theta}^{(i)} + \mathbf{A}^T \left(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}^{(i)} \right) \right] \tag{35}$$

which is the form used in [7]. Parameters α and β can be adjusted to maximize the speed of the SOSIM (see [7]).

In summary, the resulting method is a GMM algorithm where each step consists of computing matrix $\mathbf{E}$, followed by a number of SOSIM steps, large enough to guarantee the decrease of the Q-function.

D. Singularities and Convergence

The main difficulty in studying convergence of the algorithm defined by (29) and (30) is caused by the following feature: if a component reaches zero, it stays zero forever [see (21)], possibly preventing convergence to a minimizer of the objective function.

A similar difficulty appears in the IRLS algorithm for robust regression and has caused serious problems in characterizing its convergence behavior; e.g., the convergence proof in [11] includes a finiteness condition on the weights which, in our problem, would require using a penalty function with second derivative at the origin. As noted above, this would rule out most sparseness inducing penalties, which are not differentiable at the origin.

A related issue occurs in the so-called Weiszfeld algorithm (WA) [52] for the Fermat–Weber problem, which consists in finding the point minimizing the sum of the distances to a set of given points (see [9] for recent results and references). The WA can also be seen as an MM algorithm based on quadratic majorization and also has an IRLS flavor [12]. The proof of convergence of that algorithm requires that all weights are always finite, and most of the work thereafter was focused on studying conditions under which this is true.

The observations in the previous paragraph clearly beg the following question: if the algorithm is initialized with all components different from zero, does it converge to a minimizer of the objective function? Although we do not have a proof of convergence, we will next present results (the proofs of which can be found in Appendix A) which strongly suggest that this IRLS-type zero locking behavior does not seem to compromise the convergence of the algorithm.

Definition 1: Let $Z(\boldsymbol{\theta}) = \{i : \theta_i = 0\}$ and $\widetilde{Z}(\boldsymbol{\theta}) = \{i : \theta_i \neq 0\}$ be two functions that return the sets of indices of the, respectively, zero and nonzero components of a vector.

Proposition 2: Consider that $\mathbf{y}$ is generated according to (2), i.e., $\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{n}$, with $\boldsymbol{\theta}$ an arbitrary fixed-parameter vector, and the update equation is given by (32). Then

$$Z(\boldsymbol{\theta}') = \emptyset \Rightarrow \mathbb{P}\left(\{\mathbf{y} : Z(\boldsymbol{\theta}'') \neq \emptyset\} \right) = 0 \tag{36}$$

that is, with probability one with respect to the (Gaussian) density governing the generation of $\mathbf{y}$, if the algorithm is initialized such that $Z(\widehat{\boldsymbol{\theta}}^{(1)}) = \emptyset$, then, $Z(\widehat{\boldsymbol{\theta}}^{(t)}) = \emptyset$, for any finite t.

The following proposition characterizes the minima of the objective function (5) and extends to arbitrary convex GSM priors recent results shown in [30] and [31] for $C(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_1$.

Proposition 3: Consider the objective function $L(\boldsymbol{\theta}) = ||\mathbf{y} - \mathbf{A}\boldsymbol{\theta}||^2/2 + \lambda C(\boldsymbol{\theta})$ where $C(\boldsymbol{\theta}) = \sum_i C(\theta_i)$ is a sum of convex (not necessarily strictly so) even functions, continuously differentiable everywhere except maybe at the origin. Then, $\boldsymbol{\theta}^*$ is a global minimum of $L(\boldsymbol{\theta})$ if and only if its components satisfy the following set of conditions:

$$\mathbf{a}_j^T(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}^*) = \lambda C'\left(\theta_j^*\right), \quad \text{if} \quad \theta_j^* \neq 0 \tag{37}$$

$$\left|\mathbf{a}_j^T\left(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}^*\right)\right| \leq \lambda\delta, \quad \text{if} \quad \theta_j^* = 0 \tag{38}$$

where $\mathbf{a}_j$ is the jth column of matrix $\mathbf{A}$ and $\delta = C'(0^+) \equiv \lim_{\theta \to 0^+} C'(\theta)$.

Finally, the following proposition uses the previous one to characterize the points to which the algorithm may converge.

Proposition 4: Let the iterative algorithm defined by the update (32) be initialized with all nonzero components, i.e., $Z(\widehat{\boldsymbol{\theta}}^{(1)}) = \emptyset$. If the algorithm converges to some point $\boldsymbol{\theta}^{\times}$, then, with probability one, this point satisfies the necessary and sufficient conditions (NSC) of optimality (37), (38), thus, is a global optimum.

In summary, we have shown that if the algorithm is initialized with all components different from zero, then (with probability one) no component will become zero in a finite number of steps; moreover, if the algorithm converges, then (also with probability one) it does so to a global optimum of the convex objective function. Notice that these results say nothing about rates of convergence, and it is not clear how the proximity of singularities affects the speed of the algorithm; this is left as a topic of future research.

VI. MM Algorithms by Majorizing Both the Log-Likelihood and the Penalty

A. Quadratic Majorizers

It is clear from **Property 2** (see Section III) that a third class of MM algorithms can be obtained by combining (i.e., adding) the majorizers (13) and (25) derived in the two previous sections, yielding the Q-function

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}') = ||\boldsymbol{\theta} - \boldsymbol{\phi}||_2^2 + \sum_i q\left(\theta_i; \theta_i'\right). \tag{39}$$

Notice that $Q(\boldsymbol{\theta}; \boldsymbol{\theta}')$ can be minimized separately w.r.t. each component θ_i, leading to a simple shrinkage operation,

$$\theta_i'' = \frac{E_{i,i}\phi_i}{1 + E_{i,i}} \tag{40}$$

where $E_{i,i}$ depends on the previous estimate according to (33). Observe [see (35)] that this update rule coincides with a single SOSIM iteration for $\alpha = \beta = 1$ (with $\alpha = 1$, the SOSIM is in fact a first-order method).

B. Nonquadratic Majorizer for the Penalty

The fact that the majorizer on the log-likelihood makes this term separable opens the door to the use of majorizers on the penalty which need not be quadratic. In fact, what is desirable is that the penalty majorizer, when added to a separable log-likelihood majorizer, yields a Q-function with a closed-form minimizer. In view of this, an ℓ_1 majorizer is a natural choice for ℓ_p penalties with $0 < p < 1$, for two reasons: it is tighter than a quadratic majorizer; the minimizer of the resulting Qfunction is given by a simple soft thresholding rule.

The penalty $|\theta|^p$, for $0 < p < 1$ and $\theta' \neq 0$, satisfies the inequality

$$|\theta|^p \leq |\theta|p|\theta'|^{p-1} + (1-p)|\theta'|^p \tag{41}$$

with equality for $\theta = \theta'$. Of course, for $p < 1$, the majorizer (41) is undefined for $\theta' = 0$. Proceeding as for the quadratic majorizer, we define the function $r(\cdot,\cdot) : \mathbb{R}^2 \to \overline{\mathbb{R}}$ as

$$r(\theta; \theta') = \begin{cases} \Gamma(\theta')|\theta| + \zeta, & \Leftarrow \quad \theta' \neq 0 \\ +\infty, & \Leftarrow \quad \theta' = 0 \wedge \theta \neq 0 \\ 0, & \Leftarrow \quad \theta' = 0 \wedge \theta = 0 \end{cases} \tag{42}$$

where $\Gamma(\theta') \equiv \lambda p|\theta'|^{p-1}$, while $\zeta = (1-p)|\theta'|^p$ is a constant irrelevant for the resulting algorithm. Using **Property 2**, we finally have the following bound for a GGD penalty

$$\lambda||\boldsymbol{\theta}||_p^p \leq \sum_i r\left(\theta_i, \theta_i'\right). \tag{43}$$

Combining (43) with the majorizer in (13) finally leads to the Qfunction

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}') = \frac{1}{2}||\boldsymbol{\theta} - \boldsymbol{\phi}||_2^2 + \sum_i r\left(\theta_i, \theta_i'\right). \tag{44}$$

Minimizing with respect to each θ_i, leads to the update rule

$$\theta_i'' = \operatorname{soft}\left(\phi_i, \overline{\Gamma}\left(\theta_i'\right)\right) \tag{45}$$

where

$$\overline{\Gamma}\left(\theta_i'\right) = \begin{cases} +\infty, & \Leftarrow \quad \theta_i' = 0 \\ \Gamma\left(\theta_i'\right), & \Leftarrow \quad \theta_i' \neq 0. \end{cases} \tag{46}$$

Notice that $\operatorname{soft}(x, +\infty) = 0$, for any x.

As with the quadratic penalty majorizer, if a component becomes zero, it will be stuck at zero forever, which may prevent convergence to a minimizer. It is not possible to extend to this majorizer the results presented in Section V-D for the quadratic majorizer. Furthermore, notice that when $p < 1$, the objective function is nonconvex; thus, no monotonic algorithm can be guaranteed to converge to a global optimum. Nevertheless, in practice, we have never observed any convergence problems: as long as all components are initialized far away enough from zero, the algorithm always yields high quality image restorations.

TABLE I
SUMMARY OF THE ALGORITHMS: FOR EACH ALGORITHM, THE COMPUTATIONS INVOLVED IN EACH ITERATION ARE SHOWN

IST	$\widehat{\boldsymbol{\theta}}^{(t+1)} = \boldsymbol{\Psi}_{C,\lambda}\left(\boldsymbol{\Phi}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)\right)$
IRS-1	Compute $\mathbf{E}$ by equation (33) and $\mathbf{F} = \mathbf{E}\left[\mathbf{E}+\mathbf{I}\right]^{-1}$; $\widehat{\boldsymbol{\theta}}^{(t+1)} = \mathbf{F}\boldsymbol{\Phi}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)$
IRS-2	If t is multiple of M, compute $\mathbf{F}$ as in IRS-1; $\widehat{\boldsymbol{\theta}}^{(t+1)} = (\alpha-\beta)\,\widehat{\boldsymbol{\theta}}^{(t)} + (1-\alpha)\,\widehat{\boldsymbol{\theta}}^{(t-1)} + \beta\mathbf{F}\,\boldsymbol{\Phi}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)$
ISoft	Compute $\boldsymbol{\gamma} = [\overline{\Gamma}(\theta_1^{(t)}), ..., \overline{\Gamma}(\theta_i^{(t)}), ...]$; $\widehat{\boldsymbol{\theta}}^{(t+1)} = \text{soft}\left(\boldsymbol{\Phi}\left(\boldsymbol{\theta}^{(t)}\right), \boldsymbol{\gamma}\right)$

VII. SUMMARY OF ALGORITHMS AND COMPUTATIONAL COST ANALYSIS

In this section, we briefly summarize all the algorithms presented in this paper. The algorithm presented in Section IV (17) is called *iterative shrinkage-thresholding* (IST), since it proceeds by iteratively applying a nonlinear shrinkage-thresholding function $\Psi_{C,\lambda}$. The class of algorithms defined in Section V are termed *iteratively reweighted shrinkage* (IRS), because (32) can be seen as a shrinkage operation, in which the shrinkage weights in $\mathbf{E}$ are updated at each iteration. When a *second-order stationary iterative method* (SOSIM), defined in (35), is used to solve (31), we refer to the resulting algorithm as IRS-2. When we take a single step of a first-order method to solve (31), the resulting update equation is given by (40) and the corresponding algorithm is called IRS-1. Finally, the algorithm introduced in Section VI-B, defined by (45), is designated as ISoft (standing for *iterative soft* thresholding).

It worth pointing out that all the algorithms involve computing $\boldsymbol{\phi}^{(t)}$, as given by (15), which is nothing more than the current estimate $\widehat{\boldsymbol{\theta}}^{(t)}$ minus the gradient of the log-likelihood term. Defining the function

$$\boldsymbol{\Phi}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \mathbf{W}^T\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{W}\boldsymbol{\theta}) \tag{47}$$

we can write $\boldsymbol{\phi}^{(t)} = \boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}}^{(t)})$. With this function in hand, we summarize the algorithms considered in this paper in Table I.

In each iteration, the costs of computing $\boldsymbol{\Psi}_{C,\lambda}$ in IST, the vector additions, the diagonal product and inversion $\mathbf{E}[\mathbf{E}+\mathbf{I}]^{-1}$ in IRS-1 and IRS-2, all the multiplications by scalars and sums in IRS-2, and the soft threshold function in ISoft, are all $O(N)$, i.e., they grow linearly with the dimension of $\boldsymbol{\theta}$. Therefore, the leading term of the cost per iteration of all the algorithms comes from computing $\boldsymbol{\Phi}$. The multiplications by $\mathbf{H}$ and $\mathbf{H}^T$, in (47), can be done efficiently via FFT, with $O(N\log N)$ cost, since these matrices represent convolutions. For the multiplications by $\mathbf{W}$ and $\mathbf{W}^T$, when these matrices correspond to orthogonal or redundant wavelet bases, there are efficient algorithms with $O(N)$ and $O(N\log N)$ cost, respectively [41]. Consequently, the global cost per iteration of all the algorithms is $O(N\log N)$.

TABLE II
EXPERIMENTAL SETTING

Experiment	image	blur kernel	$C(\boldsymbol{\theta})$	BSNR
1	Cameraman	9×9 uniform	$\|\boldsymbol{\theta}\|_1$	40 dB
2	Lena	$\frac{[1,4,6,4,1]^T[1,4,6,4,1]}{256}$	$\|\boldsymbol{\theta}\|_1$	17 dB
3	Cameraman	$[1]$ (no blur)	$\|\boldsymbol{\theta}\|_{0.5}^{0.5}$	10 dB

VIII. EXPERIMENTS

The goal of the experiments reported in this section is not to assess the performance of the image restoration criteria of the form (4). This has been carried out in several other publications, in comparison with other state of the art criteria, namely in [7], [24], [28], [29], [33], and [37]. In those papers, the reader can also find examples where the visual quality of the restored images may be assessed. It is clear that the performance of such criteria (e.g., in terms of SNR improvement) does not depend on the optimization algorithm used to implement it, but only on the type of wavelets and of the penalty $C(\boldsymbol{\theta})$. On the other hand, the relative convergence speed the algorithms is essentially independent of these choices. In this paper, we use GGD priors, i.e., $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_p^p$, and simple Haar wavelets. We are well aware that this does not lead to state-of-the-art performance in terms of SNR improvement; however, the conclusions obtained concerning the relative speed of the algorithms are valid for other wavelets and penalty functions.

The experiments reported in this section were designed to evaluate the algorithms considered in this paper in three typical image restoration scenarios: strong blur with low noise (experiment 1), mild blur with medium noise (experiment 2), and no blur with strong noise (experiment 3). The details of each of these scenarios are shown in Table II. All the algorithms were initialized with all $\widehat{\theta}_i$ equal to a small constant (notice that this does not correspond to a constant image) and parameter λ was hand tuned for the best SNR improvement.

Experiment 1: In this case we consider a strong blur, corresponding to a very ill-conditioned matrix $\mathbf{H}$. The objective function $L(\boldsymbol{\theta}^{(t)})$ is plotted in Fig. 1. IRS-2 is clearly faster than IRS-1 and IST: IRS-1 and IST require roughly 3700 iterations to reach the objective function values that IRS-2 reaches after 300 iterations. This was already illustrated in [7] and is due to the ability of the SOSIM to handle ill-conditioned systems. The slowness of IST in this problem can be traced to the matrix bound in (12), with $\mathbf{D} = \mathbf{I}$, which is very loose because $\mathbf{H}$ is very ill-conditioned. In this problem, ISoft coincides with IST, because the penalty is $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$. In conclusion, of the algorithms described in this paper, IRS-2 should be chosen for problems involving severely ill-conditioned blurs.

Experiment 2: This experiment is targeted at assessing the behavior of the algorithms for mild blur and medium noise. The evolution of the objective function (in Fig. 2) shows that IST is faster than both IRS-1 and IRS-2. This is again a understandable result: with mild blur and medium noise, the problem is closer to denoising than to deblurring, and IST takes advantage of the fact that, in each iteration, it uses an exact denoising rule. Again, in this case, ISoft coincides with IST, because the adopted penalty

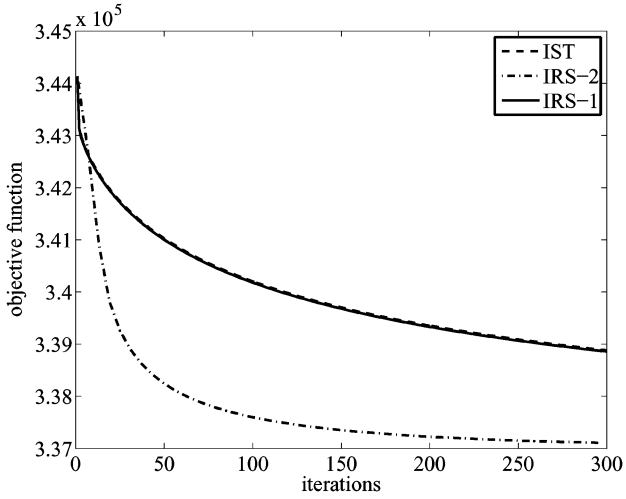

Fig. 1. Evolution of the objective function $L(\widehat{\boldsymbol{\theta}}^{(t)})$ produced by the algorithms IST, IRS-1, and IRS-2 in experiment 1 (see text and Table II for details).

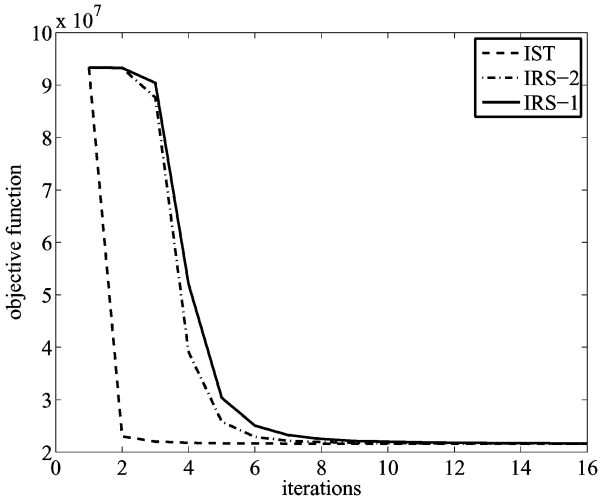

Fig. 2. Evolution of the objective function $L(\widehat{\boldsymbol{\theta}}^{(t)})$ produced by the algorithms IST, IRS-1, and IRS-2 in experiment 2 (see text and Table II for details).

is $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$. In conclusion, in problems involving mild blur and medium to strong noise, IST should be the chosen method.

Experiment 3: Finally, the third experiment aims at assessing the speed of the ISoft algorithm. Because ISoft only differs from IRS-1 and IST in the way it handles the penalty (not the likelihood), we consider a simple denoising problem, i.e., with $\mathbf{H} = \mathbf{I}$, with the penalty $C(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_p^p$ with $p = 1/2$. Notice that, in this case, the denoising rule $\boldsymbol{\Psi}_{C,\lambda}$ [see (16), and (17)] of IST does not have a closed-form; thus, we have implemented $\boldsymbol{\Psi}_{C,\lambda}$ via a numerical solution of (16). Of course, each iteration of the resulting IST scheme is computationally much heavier than each iteration of ISoft or ISR-1. Given the absence of blur, and the fact that we are using orthogonal wavelets, $\alpha = \beta = 1$ is the optimal parametrization of IRS-2, making it similar to IRS-1. The results in Fig. 3 show that ISoft is almost as fast as IST (which converges in one iteration, because this is a denoising problem) without involving the expensive numerical implementation of $\boldsymbol{\Psi}_{C,\lambda}$. ISoft is faster than IRS-1 because the quadratic bound used by the latter algorithm is not as tight as the ℓ_1 majorizer used by ISoft.

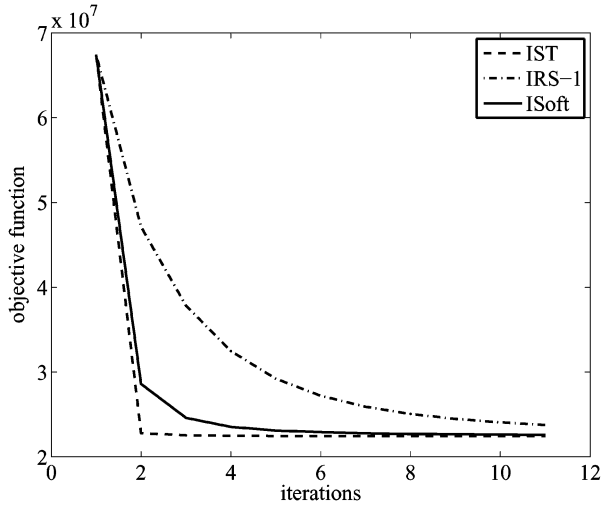

Fig. 3. Evolution of the objective function $L(\widehat{\boldsymbol{\theta}}^{(t)})$ produced by the algorithms IST, IRS-1, and ISoft in experiment 3 (see text and Table II for details).

IX. Concluding Remarks

In this paper, we have shown that several recently proposed algorithms for wavelet-based image deconvolution can all be seen as members of the MM family, resulting from different choices of majorizers. The IST class of algorithms (recently proposed by several authors) results from bounding the Hessian of the log-likelihood term with an identity matrix.

By using a quadratic majorizer on the penalty function, we obtain IRS methods. This class is further divided into IRS-1 and IRS-2, when first- or second-order iterative algorithms, respectively, are used to address the linear system that needs to be solved at each iteration. These algorithms share some features with the IRLS family, namely in that both involve weights which, in principle, and if handled naïvely, can become infinite if some component(s) of the iterate becomes zero. Moreover, once a component becomes zero, it remains there forever, possible compromising the convergence of the algorithm to a minimizer of the objective function. We have shown several results which strongly suggest that this feature of IRS algorithms does not destroy their usefulness: if properly initialized, the algorithm never (i.e., with probability zero) produces zeros in a finite number of steps; if the algorithm converges, then it does so to a minimum of the objective function. We have also shown how to write the algorithm in such a way that, even if some components become zero, no infinite weights have to be handled.

Finally, we have introduced a new class of methods, obtained by combining a bound on the log-likelihood with an ℓ_1 majorizer on the penalty. For nonconvex penalties, the ℓ_1 majorizer is tighter than the quadratic one, leading to faster algorithms.

We have experimentally compared these algorithms in typical image restoration benchmark scenarios. The conclusions of this comparison can be summarized as follows: algorithm IRS-2 is the best for problems involving severe blurs; in problems involving mild blur and medium to large noise, IST outperforms the other methods; in problems with GGD priors with exponent less than one, ISoft performs better than IRS, while IST can not be directly used because the necessary denoising rule does not have a closed-form expression.

Current research work is aimed at obtaining methods which perform as well as IRS-2 under strong blur and as well as IST in weak blur and medium to high noise situations.

APPENDIX A
PROOFS

1) Proof of Proposition 1:

Proof: The spectral norm of a symmetric matrix $\mathbf{B}$, denoted $\|\mathbf{B}\|_2$, is its largest absolute eigenvalue. If $\{\epsilon_i\}$ are the eigenvalues of $\mathbf{B}$, the eigenvalues of $\mathbf{I}-\mathbf{B}$ are $\{1-\epsilon_i\}$, thus $\|\mathbf{B}\|_2 \leq 1$ implies that $\mathbf{I} \succeq \mathbf{B}$. It turns out that

$$\begin{aligned}\|\mathbf{W}^T\mathbf{H}^T\mathbf{H}\mathbf{W}\|_2 &= \left\|\mathbf{H}\mathbf{W}(\mathbf{H}\mathbf{W})^T\right\|_2 \\ &= \|\mathbf{H}\mathbf{W}\mathbf{W}^T\mathbf{H}^T\|_2 \\ &= \|\mathbf{H}\|_2^2 = 1 \end{aligned} \tag{48}$$

because, by hypothesis, the convolution operator is normalized, i.e., $\|\mathbf{H}\|_2^2 = 1$; by hypothesis, the columns of matrix $\mathbf{W}$ correspond to a normalized tight frame, i.e., $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, [10], [41]; for any matrix $\mathbf{B}$, $\|\mathbf{B}\mathbf{B}^T\|_2 = \|\mathbf{B}^T\mathbf{B}\|_2$. ■

2) Proof of Equation (32):

Proof: Applying the matrix inversion lemma to (31), as well as the fact that all elements of $\mathbf{D}_{\widetilde{Z}}$ are nonzero

$$\begin{aligned}\boldsymbol{\theta}''_{\widetilde{Z}} &= \left(\mathbf{A}_{\widetilde{Z}}^T\mathbf{A}_{\widetilde{Z}} + \mathbf{D}_{\widetilde{Z}}\right)^{-1}\mathbf{A}_{\widetilde{Z}}^T\mathbf{y} \\ &= \left[\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T - \mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T\left(\mathbf{A}_{\widetilde{Z}}\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T + \mathbf{I}\right)^{-1}\right. \\ &\quad \left.\times\mathbf{A}_{\widetilde{Z}}\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T\right]\mathbf{y}.\end{aligned}$$

Putting the factor $\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T$ in evidence on the left, and adding and subtracting $(\mathbf{A}_{\widetilde{Z}}\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T + \mathbf{I})$ inside the square brackets

$$\begin{aligned}\boldsymbol{\theta}''_{\widetilde{Z}} &= \mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T\left[\mathbf{I} - \left(\mathbf{A}_{\widetilde{Z}}\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T + \mathbf{I}\right)^{-1}\left(\mathbf{A}_{\widetilde{Z}}\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T + \mathbf{I}\right)\right. \\ &\quad \left. + \left(\mathbf{A}_{\widetilde{Z}}\mathbf{D}_{\widetilde{Z}}^{-1}\mathbf{A}_{\widetilde{Z}}^T + \mathbf{I}\right)^{-1}\right]\mathbf{y} \\ &= \mathbf{E}_{\widetilde{Z}}\mathbf{A}_{\widetilde{Z}}^T\left(\mathbf{A}_{\widetilde{Z}}\mathbf{E}_{\widetilde{Z}}\mathbf{A}_{\widetilde{Z}}^T + \mathbf{I}\right)^{-1}\mathbf{y}\end{aligned} \tag{49}$$

where $\mathbf{E}_{\widetilde{Z}} = \mathbf{D}_{\widetilde{Z}}^{-1}$ is a diagonal matrix. Notice now that matrix $\mathbf{E}_{\widetilde{Z}}$ is simply obtained from $\mathbf{E}$ [defined in (33)] by keeping only the nonzero elements; thus

$$\mathbf{A}_{\widetilde{Z}}\mathbf{E}_{\widetilde{Z}}\mathbf{A}_{\widetilde{Z}}^T = \mathbf{A}\mathbf{E}\mathbf{A}^T.$$

Finally, it is clear that combining $\boldsymbol{\theta}''_Z = \mathbf{0}$ and the definition of $\boldsymbol{\theta}''_{\widetilde{Z}}$ given by (49) into a single equation yields (32). ■

3) Proof of Proposition 2:

Proof: Without loss of generality, consider one particular component of $\boldsymbol{\theta}''$, say θ''_j. Since all diagonal elements of $\mathbf{E}$ are nonzero (because, by hypothesis, $Z(\boldsymbol{\theta}') = \emptyset$) for θ''_j to be zero it is necessary that

$$\mathbf{a}_j^T(\mathbf{A}\mathbf{E}\mathbf{A}^T + \mathbf{I})^{-1}\mathbf{y} = 0 \tag{50}$$

where $\mathbf{a}_j$ denotes the jth column of matrix $\mathbf{A}$. This condition means that the vector $(\mathbf{A}\mathbf{E}\mathbf{A}^T + \mathbf{I})^{-1}\mathbf{y}$ must belong to the subspace orthogonal to $\mathbf{a}_j$. But matrix $(\mathbf{A}\mathbf{E}\mathbf{A}^T + \mathbf{I})$ is positive definite (because $\mathbf{A}\mathbf{E}\mathbf{A}^T$ is positive semi-definite), so it maps a subspace into a subspace, meaning that the condition in (50) is equivalent to $\mathbf{y}$ belonging to some subspace, which has zero measure, thus zero probability under the Gaussian density assumed in (4). Finally, this conclusion can be extended to the complete vector $\boldsymbol{\theta}''$, and to any finite number of iterations, since any finite union of subspaces has zero measure. ■

4) Proof of Proposition 3:

Proof: Recall that the subgradient,[5] at $\mathbf{x}$, of a convex function $f : \mathbb{R}^n \to \mathbb{R}$, denoted as $\partial f(\mathbf{x})$, is a set of vectors defined by

$$\mathbf{v} \in \partial f(\mathbf{x}) \Leftrightarrow f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{v}^T(\mathbf{y}-\mathbf{x}), \forall y \in \mathbb{R}^n.$$

If f is differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. A necessary and sufficient condition (NSC) for $L(\boldsymbol{\theta})$ to have a global minimum at $\boldsymbol{\theta}^*$ is for zero to belong to the subgradient at $\boldsymbol{\theta}^*$, i.e.,

$$\mathbf{0} \in \partial L(\boldsymbol{\theta}^*) \Leftrightarrow L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}^*), \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*. \tag{51}$$

For our objective function

$$\partial L(\boldsymbol{\theta}) = -\mathbf{A}^T(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) + \lambda\sum_i \partial C(\theta_i)$$

thus, the NSC in (51) can be written in a coordinate-wise manner as

$$\exists u_j \in \partial C\left(\theta_j^*\right) : \mathbf{a}_j^T(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) - \lambda u_j = 0, \quad \text{for all } j. \tag{52}$$

For those coordinates $\theta_j^* \neq 0$, since away from the origin $C(\theta)$ is continuously differentiable, we have $\partial C(\theta_j^*) = \{C'(\theta_j^*)\}$ and the NSC condition have the form (37).

The subgradient at zero is $\partial C(0) = [-\delta, \delta]$; this is true both if $C(\theta)$ is differentiable at the origin, in which case $\delta = 0$, or otherwise, because since $C(\theta)$ is an even function $\lim_{\theta\to 0^-} C(\theta) = -\lim_{\theta\to 0^+} C(\theta) = -\delta$. Thus, for zero coordinates, $\theta_j^* = 0$, (52) can be written as in (38). ■

5) Proof of Proposition 4:

Proof: From Proposition 2, with probability one, $Z(\widehat{\boldsymbol{\theta}}^{(t)}) = \emptyset$, for any finite t. Under this condition, $\mathbf{A}_{\widetilde{Z}} = \mathbf{A}$ and (31) can be written as

$$(\mathbf{A}^T\mathbf{A} + \mathbf{D})\widehat{\boldsymbol{\theta}}^{(t+1)} = \mathbf{A}^T\mathbf{y}. \tag{53}$$

Since $\mathbf{D}$ is diagonal and $D_{i,i} = \lambda C'(\widehat{\theta}_i^{(t)})/\widehat{\theta}_i^{(t)}$, (53) is equivalent to

$$\lambda\frac{C'\left(\widehat{\theta}_i^{(t)}\right)}{\widehat{\theta}_i^{(t)}}\widehat{\theta}_i^{(t+1)} = \mathbf{a}_i^T\left(\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{(t+1)}\right), \quad \text{for all } i. \tag{54}$$

If the algorithm converges to $\boldsymbol{\theta}^\times$, the nonzero components of $\boldsymbol{\theta}^\times$ must be fixed points of (54). Inserting this fixed-point condition

[5] See [34] for a comprehensive coverage of convex analysis.

$\widehat{\theta}_i^{(t+1)} = \widehat{\theta}_i^{(t)} = \theta_i^{\times}$ (for $\theta_i^{\times} \neq 0$) into (54) shows that these components satisfy the NSC (37).

For components that converge to zero, $\theta_i^{\times} = 0$, a fixed-point argument cannot be used, because zero components are necessarily fixed by construction of the algorithm [see (27)]. For these components, we have to explicitly study the conditions under which $\lim_{t\to\infty} \widehat{\theta}_i^{(t)} = 0$. Given that $\widehat{\theta}_i^{(t)}$ is different from zero, we can rewrite the update equation (54), as

$$\widehat{\theta}_i^{(t+1)} = \widehat{\theta}_i^{(t)} \underbrace{\left(\frac{\mathbf{a}_i^T \left(\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{(t)} \right)}{C'\left(\widehat{\theta}_i^{(t)}\right)\lambda} \right)}_{T\left(\widehat{\theta}_i^{(t)}\right)}. \tag{55}$$

Under the hypothesis that $\lim_{t\to\infty} \widehat{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{\times}$, then $|T(\widehat{\theta}_i^{(t)})|$ converges in $\overline{\mathbb{R}}$: in fact, the numerator converges to some finite number $\mathbf{a}_i^T(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}^{\times})$ and $\lambda|C'(\widehat{\theta}_i^{(t)})|$ converges to $\delta\lambda$ (recall that $\delta = \lim_{\theta\to 0^+} C'(\theta)$). If $\delta > 0$, then $|T(\widehat{\theta}_i^{(t)})|$ converges to a finite quantity, while if $\delta = 0$, $|T(\widehat{\theta}_i^{(t)})|$ goes to $+\infty$. For $\theta_i^{(t)}$ to converge to zero it is, thus, necessary that $|T(\theta_i^{\times})| < 1$. Finally, notice that this condition is the same as (38). ■

REFERENCES

[1] H. Andrews and B. Hunt, *Digital Image Restoration*. Englewood Cliffs, NJ: Prentice-Hall, 1977.

[2] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Roy. Statist. Soc. B*, vol. 36, pp. 99–102, 1974.

[3] A. Antoniadis and J. Fan, "Regularized wavelet approximations," *J. Amer. Statist. Assoc.*, vol. 96, pp. 939–967, 2001.

[4] O. Axelsson, *Iterative Solution Methods*. Cambridg, U.K.: Cambridge Univ. Press, 1996.

[5] M. Banham and A. Katsaggelos, "Spatially adaptive wavelet-based multiscale image restoration," *IEEE Trans. Image Process.*, vol. 5, no. 4, pp. 619–634, Apr. 1996.

[6] M. Belge, M. E. Kilmer, and E. L. Miller, "Wavelet domain image restoration with adaptive edge-peserving regularization," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 597–608, Apr. 2000.

[7] J. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 937–951, Apr. 2006.

[8] D. Böhning and B. Lindsay, "Monotonicity of quadratic-approximation algorithms," *Ann. Inst. Statist. Math.*, vol. 40, pp. 641–663, 1988.

[9] J. Brimberg, "Further notes on the convergence of the Weiszfeld algorithm," *Yugoslav J. Oper. Res.*, vol. 13, pp. 199–206, 2003.

[10] C. Burrus, R. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[11] R. Byrd and D. Payne, "Convergence of the IRLS algorithm for robust regression," Tech. Rep. 313 John Hopkins Univ., Baltimore, MD, 1979.

[12] T. Chan and P. Mulet, "On the convergence of the lagged diffusivity fixed point method in total variation image restoration," *SIAM J. Numer. Anal.*, vol. 36, pp. 354–367, 1999.

[13] C. Chaux, P. Combettes, J.-C. Pesquet, and V. Wajs, "Iterative image deconvolution using overcomplete representations," presented at the Eur. Signal Process. Conf., Florence, Italy, 2006.

[14] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[15] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. New York: Springer-Verlag, 1995, pp. 125–150.

[16] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM J. Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.

[17] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[18] I. Daubechies, M. De Friese, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[19] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[20] A. de Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Trans. Med. Imag.*, vol. 14, no. 1, pp. 132–137, Mar. 1995.

[21] D. Donoho, "Nonlinear solution of linear inverse problems by wavelet-vaguelette decompositions," *J. Appl. Comput. Harmon. Anal.*, vol. 1, pp. 100–115, 1995.

[22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, pp. 407–449, 2004.

[23] M. Elad, "Why simple shrinkage is still relevant for redundant representations?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5559–5569, Dec. 2006.

[24] M. Elad, B. Matalon, and M. Zibulevsky, "Image denoising with shrinkage and redundant representations," presented at the IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition, New York, 2006.

[25] H. Erdoğan and J. Fessler, "Monotonic algorithms for transmission tomography," *IEEE Trans. Med. Imag.*, vol. 18, no. 9, pp. 801–814, Sep. 1999.

[26] M. Figueiredo and R. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffreys' noninformative prior," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1322–1331, Sep. 2001.

[27] M. Figueiredo and R. Nowak, "Wavelet-based adaptive image deconvolution," presented at the IEEE Int. Conf. Acoustics, Speech and Signal Processing, Orlando, FL, 2002.

[28] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[29] M. Figueiredo and R. Nowak, "A bound optimization approach to wavelet-based image deconvolution," presented at the IEEE Int. Conf. Image Processing, Genoa, Italy, 2005.

[30] J.-J. Fuchs, "More on sparse representations in arbitrary bases," in *Proc. 13th IFAC-IFORS Symp. Identification and System Parameter Estimation*, Rotterdam, The Netherlands, 2003, vol. 2, pp. 1357–1362.

[31] J.-J. Fuchs, "On sparse representations in arbitrary redundant basis," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004.

[32] F. Girosi, "Models of noise and robust estimates," Paper 66, Center for Biological and Computational Learning, Massachusetts Inst. Technol., Cambridge, 1991.

[33] J. Guerrero-Colon and J. Portilla, "Deblurring-by-denoising using spatially adaptive Gaussian scale mixtures in overcomplete pyramids," presented at the IEEE Int. Conf. Image Processing, Atlanta, GA, 2006.

[34] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*. Berlin, Germany: Springer-Verlag, 1993.

[35] P. Huber, *Robust Statistics*. New York: Wiley, 1981.

[36] D. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, pp. 30–37, 2004.

[37] A. Jalobeanu, N. Kingsbury, and J. Zerubia, "Image deconvolution using hidden Markov tree modeling of complex wavelet packets," presented at the IEEE Int. Conf. Image Processing, Thessaloniki, Greece, 2001.

[38] M. Lang, H. Guo, J. Odegard, C. Burrus, and R. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Process. Lett.*, vol. 3, no. 1, pp. 10–12, Jan. 1996.

[39] K. Lange and J. Fessler, "Globally convergent algorithms for maximum *a posteriori* transmission tomography," *IEEE Trans. Image Process.*, vol. 4, no. 10, pp. 1430–1438, Oct. 1995.

[40] J. Liu and P. Moulin, "Complexity-regularized image restoration," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, 1998, vol. 1, pp. 555–559.

[41] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.

[42] M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 300–303, Dec. 1999.

[43] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized-Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 909–919, Apr. 1999.

[44] R. Neelamani, H. Choi, and R. Baraniuk, "ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 418–433, Feb. 2004.

[45] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
[46] I. Schoenberg, "Metric spaces and completely monotonic functions," *Ann. Math.*, vol. 39, pp. 811–841, 1938.
[47] L. Sendur and I. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 438–441, Dec. 2002.
[48] E. Simoncelli and E. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. IEEE Int. Conf. Image Processing*, Lausanne, Switzerland, 1996, vol. I, pp. 379–382.
[49] J.-L. Starck, E. Candès, and D. Donoho, "Astronomical image representation by the curvelet transform," *Astron. Astrophys.*, vol. 398, pp. 785–800, 2003.
[50] J.-L. Starck, M. Nguyen, and F. Murtagh, "Wavelets and curvelets for image deconvolution: A combined approach," *Signal Process.*, vol. 83, pp. 2279–2283, 2003.
[51] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.
[52] E. Weiszfeld, "Sur le point pour lequel la somme des distances de n points donnés est minimum," *Tôhoku Math. J.*, vol. 43, pp. 355–386, 1937.
[53] C. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, pp. 95–103, 1983.

Mário A. T. Figueiredo (S'87–M'95–SM'00) received the E.E., M.Sc., Ph.D., and "Agregado" degrees in electrical and computer engineering from the Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal, in 1985, 1990, 1994, and 2004, respectively. Since 1994, he has been with the faculty of the Department of Electrical and Computer Engineering, IST.

He is a Researcher and Area Coordinator at the Institute of Telecommunications, Lisbon. He held visiting positions with the Department of Computer Science and Engineering, Michigan State University, East Lansing, and the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, in 1998 and 2005, respectively. His scientific interests include image processing and analysis, computer vision, statistical pattern recognition, and statistical learning.

Dr. Figueiredo received the Portuguese IBM Scientific Prize in 1995 for work on unsupervised image restoration. He is a member of the IEEE Image and Multidimensional Signal Processing Technical Commitee, he is/was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MOBILE COMPUTING, *Pattern Recognition Letters*, and *Signal Processing*. He was Guest Co-Editor of special issues of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He was a Co-Chair of the 2001 and 2003 Workshops on Energy Minimization Methods in Computer Vision and Pattern Recognition. He has been a member of program committees of several top international conferences, including CVPR, ECCV, ICIAR, ICASSP, ICIP, ICML, ICPR, MLSP, and NIPS.

José M. Bioucas-Dias (S'87–M'95) received the E.E., M.Sc., and Ph.D. and "Agregado" degrees in electrical and computer engineering from the Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal, in 1985, 1991, 1995, and 2007, respectively.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, IST. He is also a Researcher with the Communication Theory and Pattern Recognition Group, Institute of Telecommunications. His scientific interests include signal and image processing, pattern recognition, optimization, and remote sensing imaging.

Dr. Bioucas-Dias was an Associate Editor of the IEEE TRANSACTION ON CIRCUITS AND SYSTEMS II. He has been a member of programm committees of several international conferences, including CVPR, IGARSS, and ISVC. He is a researcher of several national and international research projects and networks including the Marie Curie Actions "Hyperspectral Imaging Network (HYPER-I-NET)" and the "European Doctoral Program in Signal Processing (SIGNAL)."

Robert D. Nowak (S'90–M'95–SM'04) received the B.S. (with highest distinction), M.S., and Ph.D. degrees in electrical engineering from the University of Wisconsin, Madison in 1990, 1992, and 1995, respectively.

He was a Postdoctoral Fellow at Rice University, Houston, TX, during 1995 and 1996, an Assistant Professor at Michigan State University, East Lansing, from 1996 to 1999, held Assistant and Associate Professor positions at Rice University from 1999 to 2003, and was a Visiting Professor at INRIA in 2001. He is now the McFarland-Bascom Professor of Engineering at the University of Wisconsin-Madison. His research interests include statistical signal processing, machine learning, imaging and network science, and applications in communications, bio/medical imaging, and genomics.

Dr. Nowak has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and is currently an Associate Editor for the ACM Transactions on Sensor Networks and the Secretary of the SIAM Activity Group on Imaging Science. He has also served as a Technical Program Chair for the IEEE Statistical Signal Processing Workshop and the IEEE/ACM International Symposium on Information Processing in Sensor Networks. He received the General Electric Genius of Invention Award in 1993, the National Science Foundation CAREER Award in 1997, the Army Research Office Young Investigator Program Award in 1999, the Office of Naval Research Young Investigator Program Award in 2000, and the IEEE Signal Processing Society Young Author Best Paper Award in 2000.