An EM Algorithm for Wavelet-Based Image Restoration

Mário A. T. Figueiredo, *Senior Member, IEEE*, and Robert D. Nowak, *Member, IEEE*

Abstract

This paper introduces an *expectation-maximization* (EM) algorithm for image restoration (deconvolution) based on a penalized likelihood formulated in the wavelet domain. Regularization is achieved by promoting a reconstruction with low-complexity, expressed in terms of the wavelet coefficients, taking advantage of the well known sparsity of wavelet representations. Previous works have investigated wavelet-based restoration but, except for certain special cases, the resulting criteria are solved approximately or require very demanding optimization methods. The EM algorithm herein proposed combines the efficient image representation offered by the discrete wavelet transform (DWT) with the diagonalization of the convolution operator obtained in the Fourier domain. The algorithm alternates between an E-step based on the fast Fourier transform (FFT) and a DWT-based M-step, resulting in an efficient iterative process requiring $O(N \log N)$ operations per iteration. Thus, it is the first image restoration algorithm that optimizes a wavelet-based penalized likelihood criterion and has computational complexity comparable to that of standard wavelet denoising or frequency domain deconvolution methods. The convergence behavior of the algorithm is investigated, and it is shown that under mild conditions the algorithm converges to a globally optimal restoration. Moreover, our new approach outperforms several of the best existing methods in benchmark tests, and in some cases is also much less computationally demanding.

M. Figueiredo is with the Institute of Telecommunications and the Department of Electrical and Computer Engineering, *Instituto Superior Técnico*, 1049-001 Lisboa, Portugal (E-mail: mtf@lx.it.pt).

R. Nowak is with the Department of Electrical and Computer Engineering, 2041 Duncan Hall, Rice University, Houston, TX, USA (e-mail: nowak@rice.edu).

M. Figueiredo was partially supported the Science and Technology Foundation (FCT, Portugal), grant POSI/33143/SRI/2000.

R. Nowak was partially supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0349, the Office of Naval Research, grant no. N00014-00-1-0390.

I. Introduction

Wavelet-based methods have had a decided impact on the field of image processing, especially in coding and denoising. Their success is due to the fact that the wavelet transforms of images tend to be sparse (i.e., most of the wavelet coefficients are close to zero). This implies that image approximations based on a small subset of wavelets are typically very accurate, which is a key to wavelet-based compression. The MSE performance of wavelet-based denoising is also intimately related to the approximation capabilities of wavelets. Thus, the conventional wisdom is that wavelet representations that provide good approximations will also perform well in estimation problems [19].

Image deconvolution is a more challenging problem than denoising. In addition to additive noise, the observed image is convolved with an undesired point spread response function associated with the imaging system. This is a classic, well-studied image processing task [1], but applying wavelets has proved to be a challenging problem. Deconvolution is most easily dealt with (at least computationally) in the Fourier domain. However, image modelling and denoising is better handled in the wavelet domain; here lies the problem. Convolution operators are generally quite difficult to represent in the wavelet domain, unlike the simple diagonalization obtained in the Fourier domain. This naturally suggests the possibility of combining Fourier-based deconvolution and wavelet-based denoising, and several ad hoc proposals for this sort of combination have appeared in the literature.

In this paper we formally develop an image deconvolution algorithm based on a *maximum penalized likelihood estimator* (MPLE). The MPLE cannot be computed in closed-form, and so we propose an *expectation-maximization* (EM) algorithm to numerically compute it. The result is an EM iterative deconvolution algorithm which alternates between the Fourier and wavelet domains. We compare our algorithm with several existing state-of-the-art methods in benchmark problems and show that it performs better.

II. Problem Formulation

The goal of *image restoration*, or *image reconstruction*, is to recover an *original image* $\mathbf{x}$ from a degraded (or imperfect) observed version $\mathbf{y}$ [1]. In this paper, $\mathbf{x}$ and $\mathbf{y}$ will denote vectors which contain all the image pixel values after some (e.g., lexicographic) ordering. Let N_x and N_y be the dimensionality of vectors $\mathbf{x}$ and $\mathbf{y}$, respectively.

The class of observation/degradation mechanisms considered in this paper is described by the standard "linear observation plus Gaussian noise" model:

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

In Equation (1), $\mathbf{H}$ denotes the (linear) observation operator (*i.e.*, a $N_y \times N_x$ matrix), and $\mathbf{n}$ is a sample of a zero-mean white Gaussian noise field with variance σ^2; that is, $p(\mathbf{n}) = \mathcal{N}(\mathbf{n}|\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, evaluated at $\mathbf{g}$, and $\mathbf{I}$ is an identity matrix. Typical observation mechanisms which are adequately approximated by Equation (1) include: optical blur, motion blur, tomographic projections, refraction and/or multipath effects (e.g., in underwater imaging), electronic noise, photoelectric noise.

More specifically, in this paper we are interested in problems where the observation operator models space-invariant (periodic) convolutions in the original image domain. This class of problems are usually termed *image deconvolution* or *image restoration.* The corresponding matrix $\mathbf{H}$ is then square (with $N_x = N_y = N$) block-circulant and can be diagonalized by the 2D discrete Fourier transform (DFT):

$$\mathbf{H} = \mathbf{U}^H\mathbf{D}\mathbf{U}. \tag{2}$$

In the above equation, $\mathbf{U}$ is the matrix that represents the 2D discrete Fourier transform, $\mathbf{U}^H = \mathbf{U}$ is its inverse (since $\mathbf{U}$ is a unitary matrix, that is, $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}$, where $(\cdot)^H$ denotes conjugate transpose), and $\mathbf{D}$ is a diagonal matrix containing the DFT coefficients of the convolution operator represented by $\mathbf{H}$. This means that multiplication by $\mathbf{H}$ can be performed in the discrete

Fourier domain with a simple point-wise multiplication (recall that $\mathbf{D}$ is diagonal)

$$\mathbf{H}\mathbf{x} = \mathbf{U}^H\mathbf{D}\mathbf{U}\mathbf{x} = \mathbf{U}^H\mathbf{D}\widetilde{\mathbf{x}},$$

where $\widetilde{\mathbf{x}} = \mathbf{U}\mathbf{x}$ denotes the DFT of $\mathbf{x}$.

III. Review of FFT-Based Recovery and Wiener Filtering

If $\mathbf{H}$ is invertible (*i.e.*, there are no zeros in the diagonal of $\mathbf{D}$, thus $\mathbf{D}^{-1}$ exists) we can write $\mathbf{H}^{-\mathbf{1}} = \mathbf{U}^H\mathbf{D}^{-\mathbf{1}}\mathbf{U}$. Then, if we ignore the presence of noise, we can obtain an estimate of $\mathbf{x}$ as

$$\widehat{\mathbf{x}} = \mathbf{U}^H\mathbf{D}^{-\mathbf{1}}\mathbf{U}\mathbf{y} = \mathbf{U}^H\mathbf{D}^{-\mathbf{1}}\,\widetilde{\mathbf{y}}, \tag{3}$$

where $\widetilde{\mathbf{y}} = \mathbf{U}\mathbf{y}$ denotes the DFT of the observation $\mathbf{y}$. Of course, in practice, the DFT and its inverse are computed by using the *fast Fourier transform* (FFT) algorithm, which requires $O(N\log N)$ operations (where $N = N_x = N_y$ is the number of pixels), and not with matrix multiplications. Consequently, implementing Equation (3) also requires $O(N\log N)$ operations.

In most cases of interest, $\mathbf{H}$ is ill-conditioned or even non-invertible (there are very small values, or even zeros, in the diagonal of $\mathbf{D}$) and direct inversion leads to a dramatic amplification of the observation noise or is even impossible. Therefore, some regularization procedure is required. A common choice is to adopt a *maximum penalized likelihood estimator* (MPLE)

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \left\{\log p(\mathbf{y}|\mathbf{x}) - pen(\mathbf{x})\right\} = \arg\min_{\mathbf{x}} \left\{-\log p(\mathbf{y}|\mathbf{x}) + pen(\mathbf{x})\right\}, \tag{4}$$

where $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{H}\mathbf{x}, \sigma^2\mathbf{I})$ is the likelihood function corresponding to the observation model in (1), and $pen(\mathbf{x})$ is a penalty function. From a Bayesian perspective, this is a *maximum a posteriori* (MAP) criterion under the prior $p(\mathbf{x})$, such that $pen(\mathbf{x}) = -\log p(\mathbf{x})$.

If the prior $p(\mathbf{x})$ is Gaussian, with mean $\boldsymbol{\mu}$ (usually zero) and covariance matrix $\mathbf{G}$, it is well-known (see, for example, [28]) that the MPLE/MAP estimate can be written as

$$\begin{aligned}
\widehat{\mathbf{x}} &= \arg\min_{\mathbf{x}} \left\{ \frac{1}{\sigma^2}\,\|\mathbf{H}\,\mathbf{x} - \mathbf{y}\|^2 + (\mathbf{x}-\boldsymbol{\mu})^H\mathbf{G}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}) \right\} \\
&= \boldsymbol{\mu} + \mathbf{G}\mathbf{H}^H\left(\sigma^2\mathbf{I} + \mathbf{H}\mathbf{G}\mathbf{H}^H\right)^{-1}(\mathbf{y} - \mathbf{H}\boldsymbol{\mu}).
\end{aligned} \tag{5}$$

When the covariance of the prior $\mathbf{G}$ is also (as the observation matrix $\mathbf{H}$) block-circulant (meaning that the original image is considered a sample of stationary Gaussian field with periodic boundary conditions), it is also diagonalized by the DFT and we can write $\mathbf{G} = \mathbf{U}^H\mathbf{C}\mathbf{U}$, where $\mathbf{C}$ is diagonal. In this case, Equation (5), can be implemented in the DFT domain as

$$\widehat{\mathbf{x}} = \boldsymbol{\mu} + \mathbf{U}^H\mathbf{C}\mathbf{D}^H\left(\sigma^2\mathbf{I} + \mathbf{D}\mathbf{C}\mathbf{D}^H\right)^{-1}(\mathbf{U}\mathbf{y} - \mathbf{D}\mathbf{U}\boldsymbol{\mu}). \tag{6}$$

Since the matrix being inverted in Equation (6) is diagonal, the leading computational cost is the $O(N \log N)$ corresponding to the FFTs $\mathbf{U}\boldsymbol{\mu}$ and $\mathbf{U}\mathbf{y}$ and to the inverse FFT expressed by the left multiplication by $\mathbf{U}^H$. Equation (6) is known as a Wiener filter [1].

Unfortunately, this FFT-based procedure only discriminates between signal and noise in the frequency domain. It is well-known that real-world images are not well modelled by stationary Gaussian fields. A typical image $\mathbf{x}$ will not admit a sparse Fourier representation; the signal energy may not be concentrated in a small subspace, making it difficult to remove noise and preserve signal simultaneously.

IV. Wavelet-Based Image Restoration

A. Introduction

In wavelet-based estimation, the image $\mathbf{x}$ is re-expressed in terms of an orthogonal wavelet expansion, which typically provides a very sparse representation (a few large coefficients and many very small ones) [19]. Let $\mathbf{W}$ denote the (inverse) discrete wavelet transform (DWT) and let us write $\mathbf{x} \equiv \mathbf{W}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ are the wavelet coefficients of $\mathbf{x}$ [19]. As above, let us consider an MPLE/MAP criterion for our recovery problem, expressed in terms of $\boldsymbol{\theta}$, the wavelet coefficients of the original image, that is, taking the likelihood function to be $p(\mathbf{y}|\boldsymbol{\theta})$. Considering some penalty $pen(\boldsymbol{\theta})$ emphasizing sparsity of the DWT coefficients, the MPLE/MAP estimate is given by

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathbf{y}|\boldsymbol{\theta}) + pen(\boldsymbol{\theta})\} && (7)\\
&= \arg\min_{\boldsymbol{\theta}} \left\{ \frac{\|\mathbf{y} - \mathbf{H}\mathbf{W}\boldsymbol{\theta}\|^2}{2\sigma^2} + pen(\boldsymbol{\theta}) \right\}. && (8)
\end{aligned}$$

The penalty function can be interpreted as minus the logarithm of some (non-Gaussian, sparseness-inducing) prior, $pen(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$, or as a complexity-based penalty [22], [23].

When $\mathbf{H} = \mathbf{I}$, that is, for direct denoising problems, wavelet-based methods are extremely efficient, thanks to the fast implementations of the DWT and to the orthogonality of $\mathbf{W}$ (that is, $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$) which allows solving (8) by applying a coefficient-wise denoising rule; moreover, these methods achieve state-of-the-art performance (see [12], [19], [22] and references therein). The very good performance of wavelet-based denoising methods can be traced back to the adequacy of the underlying priors/models of real world images.

Wavelet-based approaches have also been shown to be very effective in image restoration problems [2], [3], [7], [13], [14], [27], [18], [25], [30], [24], [31]. However, these methods face difficulties:

(a) unlike $\mathbf{H}$ alone, $\mathbf{HW}$ is not block-circulant, thus it is not diagonalized by the DFT;

(b) unlike $\mathbf{W}$ alone, $\mathbf{HW}$ is not orthogonal, thus precluding simple coefficient-wise denoising rules, in general.

B. Previous Work

In [18], [30], [3], general frameworks aimed at restoration approaches of the form of Equation (8) has been proposed. The results are promising, but the proposed algorithms are very numerically intensive. The iterative method of [24] is also similar in spirit, employing an ad hoc thresholding step within an iterative restoration algorithm. In certain exceptional cases in which the operator $\mathbf{H}$ is scale-homogeneous, and hence (approximately) diagonalized by $\mathbf{W}$, the so-called *wavelet-vaguelette* procedure developed by Donoho [7] leads to very efficient threshold restoration procedures. However, most convolution operators are not scale-invariant and thus the wavelet-vaguelette procedure is not applicable.

An adaptation of the *wavelet-vaguelette* approach, based on wavelet-packets designed to match the frequency behavior of certain convolutions, was proposed in [14]. This method was extended to a complex wavelet *hidden Markov tree* (see [5]) scheme in [13]. Although these methods are computationally fast, they are not applicable to most convolutions and, moreover, choosing the (image) basis to conform to the operator is exactly what wavelet methods set out to avoid in the first place. The wavelet packets matched to the frequency behavior of the convolution operator may not match image structure as well as a conventional wavelet basis.

Other methods for more general deconvolution problems have been proposed. In [2], the approach is to adapt the linear filtering spatially, based on an edge detection test. The algorithm presented in [25] combines Fourier domain regularization with wavelet domain thresholding. Another interesting recent method is the one in [27]. The methods of [13], [27], [25], [30] constitute the state-of-the-art.

Finally, we mention that EM and EM-type algorithms have been previously used in image restoration and reconstruction, although not in a wavelet-based formulation (*e.g.*, [9], [10], [15]).

V. The Best of Both Worlds

The approach proposed in this paper is able to use the best of the wavelet and Fourier worlds in image deconvolution problems. The speed and convenience of FFT-based linear filtering, which is well matched to the observation model, and the adequacy of wavelet-based image models.

A. An Equivalent Model and the EM Algorithm

Let us write the observation model in Equation (1) with respect to the DWT coefficients $\boldsymbol{\theta}$ (recall that $\mathbf{x} = \mathbf{W}\boldsymbol{\theta}$):

$$\mathbf{y} = \mathbf{H}\mathbf{W}\boldsymbol{\theta} + \mathbf{n}. \tag{9}$$

As mentioned above, this equation clearly shows where the difficulties come from: although $\mathbf{H}$ is diagonalized by the DFT, $\mathbf{HW}$ is not, and so FFT-based methods are not directly applicable. To overcome this problem, the first step consists in decomposing the white Gaussian noise $\mathbf{n}$ into the sum of two non-white Gaussian noises, i.e.,

$$\mathbf{n} = \alpha\,\mathbf{H}\,\mathbf{n}_1 + \mathbf{n}_2, \tag{10}$$

where $\mathbf{H}$ is the observation matrix, α is a positive parameter, and $\mathbf{n}_1$ and $\mathbf{n}_2$ are independent noises such that

$$\begin{aligned} p(\mathbf{n}_1) &= \mathcal{N}(\mathbf{n}_1|0,\mathbf{I}) \\ p(\mathbf{n}_2) &= \mathcal{N}(\mathbf{n}_2|0,\sigma^2\mathbf{I}-\alpha^2\mathbf{H}\mathbf{H}^T). \end{aligned}$$

For $(\sigma^2\mathbf{I}-\alpha^2\mathbf{H}\mathbf{H}^T)$ to be semi-positive definite, we must have $\alpha^2 \leq \sigma^2/\lambda_1$, where λ_1 is the largest eigenvalue of $\mathbf{H}\mathbf{H}^T$. With a normalized blur (total mass equal to one), we have $\lambda_1 = 1$, and the condition simplifies to $\alpha^2 \leq \sigma^2$. This noise decomposition allows writing a two-stage observation model, equivalent to the original one, but which involves a new latent (or hidden) image $\mathbf{z}$:

$$\begin{cases} \mathbf{z} = \mathbf{W}\boldsymbol{\theta} + \alpha\,\mathbf{n}_1 \\ \mathbf{y} = \mathbf{H}\,\mathbf{z} + \mathbf{n}_2. \end{cases} \tag{11}$$

Clearly, if we had $\mathbf{z}$, we would have a pure denoising problem with white noise (the first equation in (11)). This observation is the key to our approach since it suggests treating $\mathbf{z}$ as missing data and using the EM algorithm (see, *e.g.*, [6], [20]) to estimate $\boldsymbol{\theta}$.

In our formulation, $\mathbf{z}$ is the *missing data*, which, together with the *observed data* $\mathbf{y}$, constitutes the *complete data* $\mathbf{x} \equiv (\mathbf{y},\mathbf{z})$. The EM algorithm produces a sequence of estimates $\{\widehat{\boldsymbol{\theta}}^{(t)},\ t = 0,1,2,...\}$ by alternating two steps (until some stopping criterion is met):

- **E-step:** Computes the conditional expectation of the log-likelihood of the complete data, given the observed data and the current estimate $\widehat{\boldsymbol{\theta}}^{(t)}$. The result is the so-called Q-function:

$$Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(t)}) \equiv E\left[\log p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}) \mid \mathbf{y},\widehat{\boldsymbol{\theta}}^{(t)}\right]. \tag{12}$$

- **M-step:** Updates the estimate according to

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \{-Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(t)}) + pen(\boldsymbol{\theta})\}. \tag{13}$$

It is well known that each iteration of the EM algorithm is guaranteed to increase the penalized log-likelihood, that is,

$$-\log p(\mathbf{y}|\widehat{\boldsymbol{\theta}}^{(t+1)}) + pen(\widehat{\boldsymbol{\theta}}^{(t+1)}) \leq -\log p(\mathbf{y}|\widehat{\boldsymbol{\theta}}^{(t)}) + pen(\widehat{\boldsymbol{\theta}}^{(t)}).$$

Next, we derive the specific formulas for the E-step and the M-step, for our image deconvolution problem.

B. The E-Step: FFT-Based Estimation

The complete-data log-likelihood is $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})\, p(\mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{z})\, p(\mathbf{z}|\boldsymbol{\theta})$, because, conditioned on $\mathbf{z}$, $\mathbf{y}$ is independent of $\boldsymbol{\theta}$ (see Equation (11)). Since $\mathbf{z} = \mathbf{W}\boldsymbol{\theta} + \alpha\, \mathbf{n_1}$, where $\alpha\, \mathbf{n}_1$ is zero-mean with covariance $\alpha^2\, \mathbf{I}$, we simply have (dropping all terms that do not depend on $\boldsymbol{\theta}$):

$$\log p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) \propto \log p(\mathbf{z}|\boldsymbol{\theta}) \propto -\frac{\|\mathbf{W}\boldsymbol{\theta} - \mathbf{z}\|^2}{2\alpha^2} \propto -\frac{\boldsymbol{\theta}^T\mathbf{W}^T\mathbf{W}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{W}^T\mathbf{z}}{2\alpha^2}. \tag{14}$$

This shows that the complete-data log-likelihood is linear with respect to the missing data $\mathbf{z}$. Consequently, all that is required in the E-step is to compute the conditional expectation of $\mathbf{z}$, given the observed data $\mathbf{y}$ and current parameter estimate $\widehat{\boldsymbol{\theta}}^{(t)}$,

$$\widehat{\mathbf{z}}^{(t)} \equiv E[\mathbf{z}|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(t)}] = \int \mathbf{z}\, p(\mathbf{z}|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(t)})\, d\mathbf{z}, \tag{15}$$

and plug it into the complete-data log-likelihood to obtain

$$Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) \propto -\frac{\boldsymbol{\theta}^T\mathbf{W}^T\mathbf{W}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{W}^T\widehat{\mathbf{z}}^{(t)}}{2\alpha^2} \propto -\frac{\|\mathbf{W}\boldsymbol{\theta} - \widehat{\mathbf{z}}^{(t)}\|^2}{2\alpha^2}. \tag{16}$$

Since $p(\mathbf{y}|\mathbf{z})$ and $p(\mathbf{z}|\widehat{\boldsymbol{\theta}}^{(t)})$ are both Gaussian densities, $p(\mathbf{z}|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(t)}) \propto p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|\widehat{\boldsymbol{\theta}}^{(t)})$ is also Gaussian. Applying the Gauss-Markov theorem shows that that

$$\widehat{\mathbf{z}}^{(t)} = \mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)} + \frac{\alpha^2}{\sigma^2}\, \mathbf{U}^H\mathbf{D}^H(\mathbf{U}\mathbf{y} - \mathbf{D}\mathbf{U}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)}), \tag{17}$$

which can be efficiently implemented by FFT. Notice that $\widehat{\mathbf{x}}^{(t)} \equiv \mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)}$ can be seen as the current estimate of the true image $\mathbf{x}$. With this notation, and recalling that $\mathbf{U}^H\mathbf{D}^H\mathbf{U} = \mathbf{H}^T$ and that $\mathbf{U}^H\mathbf{D}^H\mathbf{D}\mathbf{U} = \mathbf{H}^T\mathbf{H}$, we can write the E-step as

$$\widehat{\mathbf{z}}^{(t)} = \widehat{\mathbf{x}}^{(t)} + \frac{\alpha^2}{\sigma^2}\, \mathbf{H}^T(\mathbf{y} - \mathbf{H}\widehat{\mathbf{x}}^{(t)}), \tag{18}$$

revealing its similarity with a Landweber iteration for solving $\mathbf{H}\mathbf{x} = \mathbf{y}$ [16], [29]. Of course this is just the E-step; the complete EM algorithm is not a Landweber algorithm.

C. M-Step: Wavelet-Based Denoising

In the M-step, the parameter estimate is updated as shown in Equation (13), where $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)})$ is as given by Equation (16) with $\widehat{\mathbf{z}}^{(t)}$ computed according to Equation (17):

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \left\{ \frac{\|\mathbf{W}\boldsymbol{\theta} - \widehat{\mathbf{z}}^{(t)}\|^2}{2\alpha^2} + pen(\boldsymbol{\theta}) \right\} \tag{19}$$

This is simply a MPLE/MAP estimate of $\boldsymbol{\theta}$, under the prior $p(\boldsymbol{\theta})$, for a "direct" observation denoising problem: we observe $\widehat{\mathbf{z}}^{(t)} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\theta}, \alpha^2\mathbf{I})$. Furthermore, because the wavelet transform is orthogonal we have $\|\mathbf{W}\boldsymbol{\theta} - \widehat{\mathbf{z}}^{(t)}\|^2 = \|\boldsymbol{\theta} - \widehat{\boldsymbol{\omega}}^{(t)}\|^2$, where $\widehat{\boldsymbol{\omega}}^{(t)} \equiv \boldsymbol{W}^T \widehat{\mathbf{z}}^{(t)}$ denotes the DWT transform of $\widehat{\mathbf{z}}^{(t)}$. Thus, the M-Step can be computed by applying the corresponding denoising rule to $\widehat{\boldsymbol{\omega}}^{(t)}$.

For example, under a i.i.d. Laplacian prior on the wavelet coefficients,

$$p(\boldsymbol{\theta}) \propto \exp\{-\tau\|\boldsymbol{\theta}\|_1\} \quad \Rightarrow \quad pen(\boldsymbol{\theta}) = \tau \sum_i |\theta_i|, \tag{20}$$

(where $\|\boldsymbol{\theta}\|_1 = \sum_i |\theta_i|$ denotes the l_1 norm), $\widehat{\boldsymbol{\theta}}^{(t+1)}$ is obtained by applying a *soft-threshold* function to $\widehat{\boldsymbol{\omega}}^{(t)}$, the wavelet coefficients of $\widehat{\mathbf{z}}^{(t)}$ [22]. More specifically, each component of $\widehat{\boldsymbol{\theta}}^{(t+1)}$ is obtained separately according to

$$\widehat{\theta}_i^{(t+1)} = \operatorname{sgn}\left(\widehat{\omega}_i^{(t)}\right)\left(|\widehat{\omega}_i^{(t)}| - \tau\,\alpha^2\right)_+ \tag{21}$$

where $(\cdot)_+$ denotes the *positive part operator*, defined as $(x)_+ = \max\{x, 0\}$, and $\operatorname{sgn}(\cdot)$ is the *sign function*, defined as $\operatorname{sgn}(x) = 1$, if $x > 0$, and $\operatorname{sgn}(x) = -1$, if $x < 0$. Other priors or complexity penalties will lead to different wavelet denoising rules in the M-Step [12], [19], [22], [23].

D. Computational Complexity

The computational complexity of the M-Step is dominated by the DWT, usually $O(N)$ for an orthogonal DWT. The computational load of the E-step is dominated by the $O(N \log N)$ cost of the FFT. The cost of the complete EM algorithm is thus $O(N \log N)$.

E. Some Comments

A very important feature of this EM algorithm is that any wavelet denoising procedure that can be interpreted as an MPLE/MAP rule can be employed in the M-Step. For example, $p(\boldsymbol{\theta})$ could correspond to a hidden Markov tree model [5] or to a locally adaptive model [21]; however, in this case, the M-step may not be as simple as a fixed nonlinear thresholding rule. We can also use the denoising rule that we have proposed in [11], [12], since although it was originally derived from an empirical-Bayes approach, we have shown that it corresponds to an MPLE/MAP estimate under a prior of a particular form [12]. This rule, which is given by

$$\widehat{\theta}_i^{(t+1)} = \frac{\left((\widehat{\omega}_i^{(t)})^2 - 3\,\alpha^2\right)_+}{\widehat{\omega}_i^{(t)}}, \tag{22}$$

has the important advantage of having no free parameters (thus requiring no tuning), yet it yields state-of-art performance.

Let $\mathcal{D}$ denote whichever denoising operation is applied to the wavelet coefficients (such as (21) or (22)), and $\mathcal{P}$ the resulting denoising procedure applied to some image $\mathbf{v}$, that is,

$$\mathcal{P}(\mathbf{v}) \equiv \mathbf{W}\mathcal{D}(\mathbf{W}^T\mathbf{v}). \tag{23}$$

With this notation, we can write compact expressions for each iteration of the EM algorithm

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \mathcal{D}\left(\widehat{\boldsymbol{\theta}}^{(t)} + \frac{\alpha^2}{\sigma^2}\,\mathbf{W}^T\,\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)})\right), \tag{24}$$

or, explicitly for the current image estimate,

$$\widehat{\mathbf{x}}^{(t+1)} = \mathcal{P}\left(\widehat{\mathbf{x}}^{(t)} + \frac{\alpha^2}{\sigma^2}\,\mathbf{H}^T(\mathbf{y} - \mathbf{H}\widehat{\mathbf{x}}^{(t)})\right), \tag{25}$$

which can be interpreted as a Landweber iteration followed by a wavelet-based denoising step.

Finally, let us summarize the several very attractive features of this approach:

- the computational complexity of each iteration is $O(N \log N)$;
- we can employ *any* orthogonal wavelet basis;
- we can employ *any* wavelet-based penalization.

VI. Extension to Unknown Noise Variance

A. Noise Adaptive Algorithm

Up to this point, we have assumed that the noise variance σ^2 is known in advance. We now present an extension of the proposed algorithm which also estimates σ^2. This is simply done by inserting an additional step in which the noise variance estimate is updated based on the current estimate of the true image $\widehat{\mathbf{x}}(t) \equiv \mathbf{W}\widehat{\boldsymbol{\theta}}(t)$. The complete algorithm is now defined by two steps:

- **EM step** (Equation (24)):

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \mathcal{D}\left(\widehat{\boldsymbol{\theta}}^{(t)} + \frac{\alpha^2}{\widehat{\sigma^2}^{(t)}}\mathbf{W}^T\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t)})\right), \tag{26}$$

- **Noise variance update:**

$$\widehat{\sigma^2}^{(t+1)} = \frac{\|\mathbf{H}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t+1)} - \mathbf{y}\|^2}{N}. \tag{27}$$

The complete algorithm is not an EM algorithm, but it is also guaranteed to increase the penalized likelihood function. To see that this is true, let us denote the penalized negative log-likelihood being minimized (which is now also a function of σ^2) as

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2) = \frac{N}{2}\log\sigma^2 + \frac{\|\mathbf{H}\mathbf{W}\boldsymbol{\theta} - \mathbf{y}\|^2}{2\sigma^2} + pen(\boldsymbol{\theta}). \tag{28}$$

Concerning the EM step, we know that $\mathcal{L}(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\sigma^2}^{(t)}) \leq \mathcal{L}(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\sigma^2}^{(t)})$, due to the monotonicity properties of the EM algorithm [20]. The noise variance updating step is simply a maximum likelihood estimate of σ^2, with the estimate of $\boldsymbol{\theta}$ fixed at $\widehat{\boldsymbol{\theta}}^{(t+1)}$,

$$\widehat{\sigma^2}^{(t+1)} = \frac{\|\mathbf{H}\mathbf{W}\widehat{\boldsymbol{\theta}}^{(t+1)} - \mathbf{y}\|^2}{N} = \arg\min_{\sigma^2}\mathcal{L}(\widehat{\boldsymbol{\theta}}^{(t+1)}, \sigma^2),$$

since $pen(\boldsymbol{\theta})$ does not depend on σ^2. Accordingly, we have $\mathcal{L}(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\sigma^2}^{(t+1)}) \leq \mathcal{L}(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\sigma^2}^{(t)})$. In conclusion, since both steps are guaranteed not to decrease the penalized log-likelihood function, so is their combination.

B. MAD Noise Estimate

Of course, instead of using these extra step in the algorithm, the noise variance may be estimated beforehand using, for example, the well known MAD (median absolute deviation) estimator [8]. In this case, the noise standard deviation estimate is obtained as the median absolute deviation of the finest level wavelet coefficients, divided by 0.6745. In the experiments we will compare this alternative with the noise adaptive version of the algorithm above described.

VII. Translation-Invariant Restoration

So far we have assumed an orthogonal DWT in our image restoration process, and this has played a key role in our M-step. In fact, recall that by invoking the orthogonality of the DWT, we could rewrite the M-step as

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{W}\boldsymbol{\theta} - \widehat{\mathbf{z}}^{(t)}\|^2 + 2\,\alpha^2\, pen(\boldsymbol{\theta}) \right\} \tag{29}$$

$$= \arg\min_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{\theta} - \widehat{\boldsymbol{\omega}}^{(t)}\|^2 + 2\,\alpha^2\, pen(\boldsymbol{\theta}) \right\} \tag{30}$$

where $\widehat{\boldsymbol{\omega}}^{(t)} \equiv \boldsymbol{W}^T \widehat{\mathbf{z}}^{(t)}$ is the DWT of $\mathbf{z}^{(\mathbf{t})}$. This means that, under penalties of the form $pen(\boldsymbol{\theta}) = \sum_i pen(\theta_i)$, we can process each component of $\widehat{\boldsymbol{\omega}}^{(t)}$ separately to obtain $\widehat{\boldsymbol{\theta}}^{(t+1)}$.

It is well known that the dyadic image partitioning underlying the orthogonal DWT can cause blocky artifacts in the processed images. Translation-invariant (TI) wavelet denoising methods can significantly reduce these artifacts and are routinely used instead of the orthogonal DWT [4], [17]. The TI-DWT is a redundant transform that computes the inner products between the image and all (circularly) translated versions of the wavelet basis functions. Working with all possible shifts of the discrete wavelet basis functions, rather than the dyadic shifts underlying the orthogonal DWT basis functions, helps to reduce blocky artifacts [4], [17].

The TI-DWT is a redundant, overcomplete transform, producing more coefficients than pixels values in the original image. Therefore the TI-DWT is non-invertible; the "inverse" TI-DWT is actually a pseudo-inverse, and this affects our EM algorithm in a significant way. If $\boldsymbol{W}$ corresponds to a TI-DWT, then it is not orthogonal and so equality (30) is no longer true and an exact M-Step is much more complicated.

Let us begin our analysis by recalling that the TI-DWT is an overcomplete transform based on N orthogonal DWTs. Each of the N DWTs is comprised of circularly shifted versions of the discrete DWT basis functions. Let $\boldsymbol{W}_0^T$ be an orthogonal DWT. Let $i = (i_1, i_2)$ denote a generic (circular) shift of i_1 lines and i_2 columns, and let $\boldsymbol{W}_i^T$ denote a DWT comprised of the discrete basis functions in $\boldsymbol{W}_0^T$ circularly shifted by i. There are N distinct *shifted* DWTs, which we denote by $\boldsymbol{W}_0^T, \ldots, \boldsymbol{W}_{N-1}^T$. In this section, we will denote the TI-DWT by

$$\boldsymbol{W}^T = \frac{1}{\sqrt{N}} \left[\boldsymbol{W}_0 \;\cdots\; \boldsymbol{W}_{N-1}\right]^T = \frac{1}{\sqrt{N}} \begin{bmatrix} \boldsymbol{W}_0^T \\ \vdots \\ \boldsymbol{W}_{N-1}^T \end{bmatrix}. \tag{31}$$

As mentioned above, the TI-DWT is not invertible, so the pseudo-inverse

$$\boldsymbol{W} = \frac{1}{\sqrt{N}} \left[\boldsymbol{W}_0 \;\cdots\; \boldsymbol{W}_{N-1}\right] \tag{32}$$

is used to transform the redundant set of coefficients back to the image space. Notice that if $\mathbf{x}$ denotes any image, then

$$\boldsymbol{W}\boldsymbol{W}^T\mathbf{x} = \frac{1}{N} \left[\boldsymbol{W}_0 \;\cdots\; \boldsymbol{W}_{N-1}\right] \begin{bmatrix} \boldsymbol{W}_0^T \\ \vdots \\ \boldsymbol{W}_{N-1}^T \end{bmatrix} \mathbf{x} = \mathbf{\frac{1}{N}} \sum_{\mathbf{i}=\mathbf{0}}^{\mathbf{N}-\mathbf{1}} \boldsymbol{W}_{\mathbf{i}}\boldsymbol{W}_{\mathbf{i}}^{\mathbf{T}}\mathbf{x} = \mathbf{x},$$

because $\boldsymbol{W}_i\boldsymbol{W}_i^T = \mathbf{I}$, and thus $\boldsymbol{W}\boldsymbol{W}^T = \mathbf{I}$. However, clearly, $\boldsymbol{W}^T\boldsymbol{W} \neq \mathbf{I}$ and thus $\boldsymbol{W}$ is not orthogonal.

When $\boldsymbol{W}$ corresponds to a TI-DWT[1], the M-Step of our EM algorithm can not be simplified as in (29) – (30). However, we show in the Appendix that if the penalty $pen(\boldsymbol{\theta})$ is convex, and has the form $pen(\boldsymbol{\theta}) = \sum_i pen(\theta_i)$, then solving (30) in place of (29) in the TI-DWT case, corresponds to a *generalized* EM (GEM) algorithm [20]. Specifically, the resulting M-step does not correspond to an exact minimization of (29), but produces a $\widehat{\boldsymbol{\theta}}^{(t+1)}$ satisfying a weaker condition:

$$-Q(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}) + pen(\widehat{\boldsymbol{\theta}}^{(t+1)}) \;\leq\; -Q(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) + pen(\widehat{\boldsymbol{\theta}}^{(t)}). \tag{33}$$

GEM algorithms possess the same basic monotonicity and convergence properties as the normal EM algorithm [20], [32]. Of course, the standard EM algorithm is a particular case of GEM.

[1] A similar complication arises if the orthogonal DWT is replaced by a biorthogonal DWT, but we will not investigate that problem here.

This demonstrates, for example, that applying a standard soft thresholding operation (which corresponds to $pen(\boldsymbol{\theta}) = \sum_i |\theta_i|$) to the TI-DWT coefficients does not increase the cost criterion and thus we are in the presence of a GEM algorithm [6], [32].

Finally, we recall that the coefficients of the TI-DWT can be efficiently computed using the so-called undecimated DWT (UDWT), which simply eliminates the down-sampling process in the filter-bank implementation of a wavelet transform [17]. This results in $N \log N$ coefficients instead of N. The TI-DWT produces N^2 coefficients in total, but only $N \log N$ values are unique because certain shifts generate the same inner products between the image and basis functions. The filter-bank implementation of the UDWT produces only the $N \log N$ unique coefficients, and requires $O(N \log N)$ operations. Thus, the computational complexity of each M-Step in the TI case is $O(N \log N)$.

VIII. Convergence Analysis of the EM Algorithm

A general, basic property of EM/GEM algorithms is that it generates a sequence of non-decreasing (penalized) likelihood values [20]. Iteration of the basic EM/GEM steps of the algorithm produces a sequence of images, and each image of the sequence has a penalized likelihood value greater than or equal to that of the image preceding it. This is a very desirable property, but several questions remain: (1) Does the sequence (of penalized likelihood values) converge to the maximum of the penalized likelihood function? (2) Does the corresponding sequence of images converge to a fixed image and is this limit (assuming it exists) unique? This section explores these issues. First, we consider the conditions under which the EM/GEM algorithm converges to a stationary point of the penalized likelihood function. Second, we investigate the convexity of the penalized negative log-likelihood function and establish conditions under which the EM/GEM algorithm will converge to a unique solution.

A. Convergence to a Stationary Point

The results in [32] guarantee that the EM/GEM algorithm converges to a stationary point (local maximum or saddle-point) of the penalized likelihood function under fairly mild conditions. Theorem 2 of [32] shows that all limit points of the EM/GEM algorithm are stationary points of the penalized likelihood function, provided that $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)})$ and $pen(\boldsymbol{\theta})$ are continuous in both $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$. This condition is easily verfied for the the expected complete-data log-likelihood $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)})$. The penalty function, $pen(\boldsymbol{\theta})$, also needs to be continuous in order to guarantee convergence to a stationary point. This precludes the use of the conventional hard-threshold function, but both the soft-threshold rule (21) and our rule in (22) correspond to continuous penalty functions (log-priors). Additionally, for the translation-invariant case, the penalty must also be convex (e.g., soft-threshold rule). To summarize, if the penalty function underlying the nonlinear shrinkage/threshold function employed in the M-Step is continuous in $\boldsymbol{\theta}$, then the EM algorithm based on an orthogonal DWT sequence converges to a stationary point of the penalized log-likelihood. In the TI-DWT case, the same convergence holds if the penalization function is both continuous and convex.

The limit points may be local maxima or saddle-points; it is difficult to guarantee convergence to a local maximum without further assumptions. Such conditions are investigated next.

B. Convergence to a Global Maximum

Let us begin by considering the case in which $\mathbf{H}$ is invertible and $\mathbf{W}$ is an orthogonal wavelet transform. Under these assumptions, the negative log-likelihood term of (8) is strictly convex in $\boldsymbol{\theta}$. Now if the penalty function is also convex (not necessarily strictly convex), then the penalized negative log-likelihood function is strictly convex in $\boldsymbol{\theta}$. For example, the log-Laplacian penalty function, leading to the soft-threshold rule, is convex in $\boldsymbol{\theta}$ (though not strictly so). Strict convexity of the penalized negative log-likelihood function implies that there is only one stationary point, the global maximum. Thus, under the continuity conditions discussed above, the EM algorithm is guaranteed to converge to the global maximum. Note that the uniqueness of the maximum point guarantees that the sequence of images produced by the EM algorithm converges to the global maximum penalized likelihood image restoration.

Next consider situations when $\mathbf{H}$ and/or $\mathbf{W}$ is not invertible. For examples, $\mathbf{H}$ is not invertible if the underlying point spread response has finite support, and the TI-DWT is not invertible ($\boldsymbol{\theta}$ cannot be uniquely recovered from $\mathbf{x}$). In such cases the negative log-likelihood term of (8) is convex, but not strictly convex, in $\boldsymbol{\theta}$. If the penalty function is also convex (but not strictly so), then the sequences of penalized log-likelihood values produced by the EM/GEM algorithms will converge their respective global maximum penalized log-likelihood values. This follows from the EM/GEM convergence results of Wu [32], since all stationary points of a convex function are global minima. However, since there may be many global minima, the EM/GEM algorithms may not converge to fixed images (they are only guaranteed to converge to their respective sets of images corresponding to global minima). If it does converge to a fixed image (this limit could depend on the initialization of the algorithm), then that image maximizes the penalized likelihood criterion.

If the penalty function is strictly convex, then the EM/GEM algorithms are guaranteed to converge to the unique maximum penalized likelihood value and a unique optimal image. This also follows from the EM/GEM convergence results [32]; the unique stationary point of a strictly convex function is the global minimum. So far, the only convex penalty function we have considered is the log-Laplacian (leading to the soft-threshold rule), but even this penalty function is not strictly convex, since its growth with the absolute value of the argument is linear. The following modification of the log-Laplacian leads to a strictly convex penalty function and a threshold rule nearly the same as the soft-threshold function, except that it is differentiable at all points. Instead of the log-Laplacian penalty, which has the form $-\log e^{-\alpha|\theta|} = \alpha\,|\theta|$, consider

$$pen(\theta) \;=\; -\log\, e^{-\eta\sqrt{\theta^2+\beta^2}} \;=\; \eta\sqrt{\theta^2+\beta^2}, \tag{34}$$

for some small number β. Notice that as $\beta \to 0$, this penalty tends to the log-Laplacian. However, for every $\alpha, \beta > 0$ this penalty is strictly convex, since

$$\frac{d^2(\eta\sqrt{\theta^2+\beta^2})}{d\theta^2} = \frac{\eta\beta^2}{(\theta^2+\beta^2)^{3/2}} > 0.$$

Moreover, the penalty (34) induces a threshold rule similar to the soft-threshold rule; the difference is that the modified soft-threshold rule makes a smooth transition across the threshold level as shown in Figure 1.

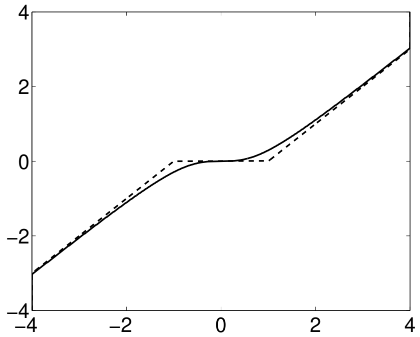

Fig. 1. Soft-threshold function (dashed) and modified soft-threshold function (solid) with threshold level set at 1 and $\beta = 1$. If $\beta = 0.1$, then the difference between the soft-threshold function and the modified soft-threshold function are indistinguishable to the naked eye at this scale.

C. Summary of Convergence Results

The following four points summarize the convergence properties of our EM/GEM algorithms.

1. If the penalty $pen(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, then each iteration of the EM/GEM algorithms produces an image with a penalized likelihood value greater than or equal to the previous image.

2. If the penalty function is also convex (but not strictly convex) in $\boldsymbol{\theta}$, then the sequence of penalized log likelihood values converges to the global maximum. However, since there may be many global maxima the EM/GEM algorithms may not converge to a fixed image. If it does converge to a fixed image, then that image maximizes the penalized likelihood criterion.

3. The EM/GEM algorithms converge to the unique, globally optimal solution (fixed image) of the penalized likelihood criterion if either of the conditions below are met:

i. $\mathbf{H}$ and $\mathbf{W}$ are invertible and the penalty function is convex (e.g., soft-threshold).

ii. The penalty function is strictly convex (e.g., the modified soft-threshold penalty (34)).

4. Recall the that EM/GEM algorithms coupled with adaptive updates of the noise variance, given by equations (26) and (27), produce non-decreasing sequences of penalized likelihood values (with the noise variance σ^2 treated as an unknown parameter to be inferred jointly with $\boldsymbol{\theta}$).

However, the corresponding penalized negative log likelihood function is highly non-convex and convergence is not guaranteed in this case.

IX. Experimental Results

In this section we present a set of experimental results illustrating the performance of the developed method in comparison with other state of the art methods recently described in [13], [25], and [18]. In all the experiments, we employ the TI-DWT (computed via UDWT filterbank method of [17]), using Daubechies-2 (Haar) wavelets. The convergence criterion used to stop the algorithm is

$$\frac{\|\hat{\mathbf{x}}^{(t+1)} - \hat{\mathbf{x}}^{(t)}\|_2}{\|\hat{\mathbf{x}}^{(t)}\|_2} < \delta$$

where δ is a threshold set to $10^{-3}\sigma^2$. In the noise-adaptive version, σ^2 is replaced by its estimate. In all the experiments reported, we use $\alpha = \sigma$; we found experimentally that this is a good general-purpose choice. The algorithm is initialized with a Wiener estimate, as given by (5), with $\boldsymbol{\mu} = 0$ and $\mathbf{G} = 10^3\mathbf{I}$.

We have restored all the blurred images using six variants of the algorithm, corresponding to two denoising functions (rule (22) and the rule corresponding to the modified Laplacian prior in equation (34)) and three noise estimation approaches: fixed known noise, adaptive algorithm described in Section VI-A, MAD noise estimate (Section VI-B). In the case of the rule corresponding to the modified Laplacian prior in equation (34) (see also Fig. 1), the parameters were $\eta = 0.35$ and $\beta = 0.02$.

In the first set of tests we replicate the experimental condition of [13]. The point spread function of the blur operator is given by $h_{ij} = (1 + i^2 + j^2)^{-1}$, for $i, j = -7, ..., 7$. Noise variances considered are $\sigma^2 = 2$ and $\sigma^2 = 8$. Figure 2 shows the original "cameraman" image, together with the two blurred and noisy images. The SNR improvements obtained are summarized in Table I, together with the results reported in [13]. The SNR improvements obtained by our method are always better than those reported in [13]; notice that [13] uses a more sophisticated wavelet transform and prior model. Finally, Figure 3 displays the restored versions of the two images ($\sigma^2 = 2$ and $\sigma^2 = 8$) obtained with rule (22) and adaptive noise variance estimation. The other restorations are not shown here since they are visually almost indistinguishable from these ones.

Fig. 2. Original image (top), blurred and noisy images (middle, $\sigma^2 = 2$; bottom, $\sigma^2 = 8$).

In a second set of tests, we consider the setup of [25]: uniform blur of size 9×9, and the noise variance is such that the SNR of the noisy image, with respect to the blurred image without noise (BSNR), is $40dB$. The SNR improvements obtained by the several algorithms are summarized in Table II. Figure 4 shows the observed and restored versions for this example.

In the final set of tests we have used the blur filter and noise variance considered in [18]. Specifically, the original image was blurred by a $5{\times}5$ separable filter with weights $[1,\ 4,\ 6,\ 4,\ 1]/16$ (in both horizontal and vertical directions) and then contaminated with white Gaussian noise

TABLE I

SNR improvements obtained by several variants of the proposed algorithm on the images shown in Figure 2

Method	$\sigma^2 = 2$	$\sigma^2 = 8$
Rule (22), adaptive noise	7.22dB	5.06dB
Rule (22), known noise	7.19dB	5.20dB
Rule (22), MAD noise estimate	7.17dB	5.16dB
Modified Laplacian, adaptive noise	6.91dB	4.88dB
Modified Laplacian, known noise	7.03dB	4.93dB
Modified Laplacian, MAD noise estimate	6.90dB	4.90dB
Results in [13]	6.75dB	4.85dB

TABLE II

SNR improvements obtained by several variants of the proposed algorithm on the blurred image shown in Figure 4

Method	
Rule (22), adaptive noise	7.02dB
Rule (22), known noise	7.56dB
Rule (22), MAD noise estimate	7.27dB
Modified Laplacian, adaptive noise	6.80dB
Modified Laplacian, known noise	7.26dB
Modified Laplacian, MAD noise estimate	7.17dB
Result in [25]	7.3dB
Result in [2]	6.7dB

Fig. 3. Restored images (left, $\sigma^2 = 2$; right, $\sigma^2 = 8$) obtained from the blurred noisy images of Figure 2.

Fig. 4. Left: blurred noisy image (9×9 uniform blur, -40dB noise); right: restored image using the noise-adaptive version of our algorithm with rule (22).

of standard deviation $\sigma = 7$. The SNR improvements obtained by the considered instances of our algorithm are reported in Table III. The original, blurred, and restored images are shown in Figure 5.

The results reported in this section allow concluding that the proposed method yields state-of-art performance on the considered experimental conditions. Another important conclusion is that using the MAD noise variance estimate is basically as good (or better) than using the noise-adaptive version of the algorithm; accordingly, due to its lower computational demand, the

TABLE III

SNR improvements obtained by several variants of the proposed algorithm on the blurred image shown in Figure 5

Method	
Rule (22), adaptive noise	3.54dB
Rule (22), known noise	3.69dB
Rule (22), MAD noise estimate	3.50dB
Modified Laplacian, adaptive noise	1.74dB
Modified Laplacian, known noise	2.31dB
Modified Laplacian, MAD noise estimate	2.11dB
Best result in [18]	1.078dB

Fig. 5. Original image (left), blurred and noisy image (midle) and restored image (right) using the noise-adaptive version of our algorithm with rule (22).

MAD estimate should be used together with the fixed σ^2 version of the algorithm.

X. Conclusions

This paper proposed a wavelet-based MPLE criterion for image deconvolution. The MPLE must be computed numerically, and we derived novel EM/GEM algorithms for this purpose. The EM/GEM algorithms lead to simple procedures that alternate between Fourier domain filtering and wavelet domain denoising to obtain optimal image restorations. Previous works have investigated related restoration criteria [2], [3], [7], [13], [14], [27], [18], [25], [30], [24], [31], but except for certain special cases [7], [14] these methods produce sub-optimal solutions or require very heavy computation. Our EM/GEM algorithms provably converge to optimal solutions and have computational complexity comparable to that of standard wavelet denoising schemes. Moreover, our approach outperforms several of the best existing methods. Our analysis sheds light on the nature of wavelet-based image restoration; in particular this is the first work we are aware of that carefully investigates and describes the subtle distinctions between the application of orthogonal and (non-orthogonal) translation-invariant DWTs to image restoration.

Appendix

In this appendix we prove the claim that by solving (30) instead of (29), in the case of the TI-DWT (i.e., with $\boldsymbol{W}^T$ being the TI-DWT defined in (31)), we obtain a GEM algorithm [20].

Let

$$L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) \equiv \; \|\widehat{\mathbf{z}}^{(t)} - \boldsymbol{W}\,\boldsymbol{\theta}\|^2 + 2\,\alpha^2\, pen(\boldsymbol{\theta})$$

denote the criterion to be minimized in the M-Step (see (29)). Next, notice that

$$||\widehat{\mathbf{z}}^{(t)} - \boldsymbol{W}\boldsymbol{\theta}||^2 \;=\; ||\widehat{\boldsymbol{\omega}}^{(t)} - \boldsymbol{\theta}||^2 - ||\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\theta} - \boldsymbol{\theta}||^2,$$

where $\widehat{\boldsymbol{\omega}}^{(t)} = \boldsymbol{W}^T\widehat{\mathbf{z}}^{(t)}$; this identity follows from $||\widehat{\boldsymbol{\omega}}^{(t)}||^2 = ||\widehat{\mathbf{z}}^{(t)T}\boldsymbol{W}\boldsymbol{W}^T\widehat{\mathbf{z}}^{(t)}|| = ||\widehat{\mathbf{z}}^{(t)}||^2$ (since $\boldsymbol{W}\boldsymbol{W}^T = \mathbf{I}$). Using this identity we can write

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = ||\widehat{\boldsymbol{\omega}}^{(t)} - \boldsymbol{\theta}||^2 + 2\,\alpha^2\, pen(\boldsymbol{\theta}) - ||\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\theta} - \boldsymbol{\theta}||^2. \tag{35}$$

To verify that $\widehat{\boldsymbol{\theta}}^{(t+1)}$ as given by (30) corresponds to the M-step of a GEM algorithm, we will prove that $L(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}) \leq L(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)})$. This proof will be done through the chain of inequalities

$$L(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}) \;\overset{(a)}{\leq}\; L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) \;\overset{(b)}{\leq}\; L(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}), \tag{36}$$

under the assumption that $pen(\boldsymbol{\theta})$ is a convex function and that it is additive in the sense that $pen(\boldsymbol{\theta}) = \sum_i pen(\theta_i)$ (*e.g.*, a squared Euclidean norm or an l_1 norm).

• **Inequality (a) in (36):** Recall that $\widehat{\boldsymbol{\theta}}^{(t+1)}$ is given by (30). We can then write

$$\begin{aligned}||\widehat{\boldsymbol{\omega}}^{(t)} - \widehat{\boldsymbol{\theta}}^{(t+1)}||^2 + 2\,\alpha^2\, pen(\widehat{\boldsymbol{\theta}}^{(t+1)}) &\leq ||\widehat{\boldsymbol{\omega}}^{(t)} - \boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}||^2 + 2\,\alpha^2\, pen(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}) \\ &= L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) + ||\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}||^2 \\ &= L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) \qquad (37)\end{aligned}$$

where the inequality results from the definition of minimum, the first equality from (35), and the second one from the fact that $\boldsymbol{W}\boldsymbol{W}^T = \mathbf{I}$. Finally, since

$$||\widehat{\boldsymbol{\omega}}^{(t)} - \widehat{\boldsymbol{\theta}}^{(t+1)}||^2 + 2\alpha^2 pen(\widehat{\boldsymbol{\theta}}^{(t+1)}) = L(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}) + ||\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t+1)} - \widehat{\boldsymbol{\theta}}^{(t+1)}||^2,$$

and $||\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t+1)} - \widehat{\boldsymbol{\theta}}^{(t+1)}||^2 \geq 0$, we have that

$$L(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}) \leq L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) - ||\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t+1)} - \widehat{\boldsymbol{\theta}}^{(t+1)}||^2 \leq L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}).$$

• **Inequality (b) in (36):** Begin by noticing that

$$\begin{aligned}L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) &= \|\widehat{\mathbf{z}}^{(t)} - \boldsymbol{W}\,\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}\|^2 + 2\,\alpha^2\, pen(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}) \\ &= \|\widehat{\mathbf{z}}^{(t)} - \boldsymbol{W}\,\widehat{\boldsymbol{\theta}}^{(t)}\|^2 + 2\,\alpha^2\, pen(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}), \qquad (38)\end{aligned}$$

(because $\boldsymbol{W}\,\boldsymbol{W}^T = \mathbf{I}$) and so

$$L(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) - L(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) = 2\,\alpha^2 \left(pen(\widehat{\boldsymbol{\theta}}^{(t)}) - pen(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)})\right). \qquad (39)$$

Next, notice that $\boldsymbol{W}^T\boldsymbol{W}$ is composed of N^2 blocks of the form $N^{-1}\boldsymbol{W}_i^T\boldsymbol{W}_j$, for $0 \leq i,j \leq N-1$, where i,j correspond to the block-row and block-column location of each block. Also notice that $\boldsymbol{W}_i = \boldsymbol{W}_j\ \mathbf{P}_{i,j}$, where $\mathbf{P}_{i,j}$ is a permutation matrix depending on i and j. Thus, $N^{-1}\boldsymbol{W}_i^T\boldsymbol{W}_j = N^{-1}\boldsymbol{W}_i^T\boldsymbol{W}_i\mathbf{P}_{i,j} = N^{-1}\mathbf{P}_{i,j}$. It then follows that each element of $\boldsymbol{W}^T\boldsymbol{W}$ is either 0 or N^{-1}, and that each row and column of $\boldsymbol{W}^T\boldsymbol{W}$ contains exactly N non-zero entries. Therefore,

$$pen(\boldsymbol{W}^T\boldsymbol{W}\widehat{\boldsymbol{\theta}}^{(t)}) = \sum_{i=1}^{N^2} pen\left(\sum_{j\in J_i}\frac{1}{N}\,\widehat{\theta}_j^{(t)}\right) \leq \sum_{i=1}^{N^2}\sum_{j\in J_i}\frac{1}{N}\,pen(\widehat{\theta}_j^{(t)}) = pen(\widehat{\boldsymbol{\theta}}^{(t)}),$$

where J_i is the set of N indices of the non-zero elements in the i-th row of $\boldsymbol{W}^T\boldsymbol{W}$, the inequality results from the convexity of $pen(\cdot)$, and the last equality from its additive nature. Combining this inequality with (39) finally leads to $L(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) \leq L(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})$.

Acknowledgments

The authors would like to thank Jeff Fessler and André Jalobeanu for their helpful discussions and insightful comments that helped to improve this paper.

References

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

[2] M. Banham and A. Katsaggelos. "Spatially adaptive wavelet-based multiscale image restoration," *IEEE Trans. on Image Processing*, vol. 5, pp. 619-634, 1996.

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

[4] R. Coifman and D. Donoho. "Translation invariant de-noising," in A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, Lecture Notes in Statistics vol. 103, pp. 125–150, Springer-Verlag, New York, 1995.

[5] M. Crouse, R. Nowak, and R. Baraniuk. "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. on Signal Processing*, vol. 46, pp. 886–902, 1998.

[6] A. Dempster, N. Laird, and D. Rubin. "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[7] D. Donoho. "Nonlinear solution of linear inverse problems by wavelet-vaguelette decompositions," *Journal of Applied and Computational Harmonic Analysis*, vol. 1, pp. 100–115, 1995.

[8] D. L. Donoho and I. M. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[9] J. Fessler, A. Hero. "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms," *IEEE Trans. on Image Processing*, vol. 4, pp. 1417-29, 1995.

[10] M. Figueiredo and J. Leitão, "Unsupervised Image Restoration and Edge Location using Compound Gauss-Markov random fields and the MDL principle", *IEEE Trans. on Image Processing*, vol. 6, pp. 1089-1102, 1997.

[11] M. Figueiredo and R. Nowak. "Bayesian wavelet-based signal estimation using non-informative priors," in *Proc. of the 32nd Asilomar Conference on Signals, Systems, and Computers*, pp. 1368 – 1373, Monterey, CA, 1998.

[12] M. Figueiredo and R. Nowak. "Wavelet-based image estimation: an empirical Bayes approach using Jeffreys' noninformative prior," *IEEE Trans. on Image Processing*, vol. 10, pp. 1322-1331, 2001.

[13] A. Jalobeanu, N. Kingsbury, and J. Zerubia. "Image deconvolution using hidden Markov tree modeling of

complex wavelet packets," in *Proc. of the IEEE Intern. Conf. on Image Proc. – ICIP'2001*, Thessaloniki, Greece, 2001.

[14] J. Kalifa and S. Mallat. "Minimax restoration and deconvolution," in P. Muller and B. Vidakovic, editors, *Bayesian Inference in Wavelet Based Models*, Springer-Verlag, New York, 1999.

[15] R. Lagendijk, J. Biemond, D. Boekee. "Identification and restoration of noisy blurred images using the expectation-maximization algorithm," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1180-1191, 1990.

[16] L. Landweber. "An iterative formula for Fredholm integral equations of the first kind," *American Journal of Mathematics*, vol. 73, pp. 615–624, 1951.

[17] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus and R. O. Wells. "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Letters*, vol. 3, pp. 10–12, 1996.

[18] J. Liu and P. Moulin. "Complexity-Regularized Image Restoration," *Proc. IEEE Int. Conf. on Image Proc. (ICIP'98)*, Vol. 1, pp. 555-559, Chicago, Oct. 1998.

[19] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, 1998.

[20] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley & Sons, New York, 1997.

[21] M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin. "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *Signal Processing Letters*, vol. 6, pp. 300–303, 1999.

[22] P. Moulin and J. Liu. "Analysis of multiresolution image denoising schemes using generalized - Gaussian and complexity priors," *IEEE Trans. on Information Theory,*, vol. 45, pp. 909–919, 1999.

[23] P. Moulin and J. Liu. "Statistical imaging and complexity regularization," *IEEE Trans. on Information Theory,*, vol. 46, pp. 1881–1895, 2000.

[24] F. Murtagh, J. Starck and A. Bijaoui, *Image Processing and Data Analysis.* Cambridge Univeristy Press, 1998.

[25] R. Neelamani, H. Choi, and R. Baraniuk. "Wavelet-based deconvolution for ill-conditioned systems," *IEEE Trans. on Image Processing*, 2001 (submitted).

[26] R. Nowak and E. Kolaczyk. "A statistical multiscale framework for Poisson inverse problems," *IEEE Trans. on Information Theory*, vol. 46, pp. 1811–1825, 2000.

[27] P. Rivaz and N. Kingsbury. "Bayesian image deconvolution and denoising using complex wavelets," in *Proc. of the IEEE Intern. Conf. on Image Proc. – ICIP'2001*, Thessaloniki, Greece, 2001.

[28] L. Scharf. *Statistical Signal Processing.* Addison Wesley Publishing Company, Reading, Massachusetts, 1991.

[29] O. Strand. "Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind," *SIAM Journal of Numerical Analysis*, vol. 11, pp. 798–825, 1974.

[30] Y. Wan and R. Nowak. "A wavelet-based approach to joint image restoration and edge detection," in *SPIE*

Conference on Wavelet Applications in Signal and Image Processing VII, Denver, CO, 1999. SPIE Vol. 3813.

[31] J. Z. G. Wang and G.-W. Pan, "Solution of inverse problem in image processing by wavelet expansion," *IEEE Trans. Image Processing*, vol. 4, no. 5, pp. 579–593, 1995.

[32] C. Wu. "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11, pp. 95-103, 1983.