Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 429128, 12 pages
doi:10.1155/2008/429128

Research Article

Iterative Estimation Algorithms Using Conjugate Function Lower Bound and Minorization-Maximization with Applications in Image Denoising

Guang Deng[1] and Wai-Yin Ng[2]

[1] *Department of Electronic Engineering, La Trobe University, Bundoora, Victoria 3086, Australia*
[2] *Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong*

Correspondence should be addressed to Guang Deng, d.deng@latrobe.edu.au

Received 19 September 2007; Revised 3 January 2008; Accepted 11 February 2008

Recommended by Hubert Cardot

A fundamental problem in signal processing is to estimate signal from noisy observations. This is usually formulated as an optimization problem. Optimizations based on variational lower bound and minorization-maximization have been widely used in machine learning research, signal processing, and statistics. In this paper, we study iterative algorithms based on the conjugate function lower bound (CFLB) and minorization-maximization (MM) for a class of objective functions. We propose a generalized version of these two algorithms and show that they are equivalent when the objective function is convex and differentiable. We then develop a CFLB/MM algorithm for solving the MAP estimation problems under a linear Gaussian observation model. We modify this algorithm for wavelet-domain image denoising. Experimental results show that using a single wavelet representation the performance of the proposed algorithms makes better than that of the bishrinkage algorithm which is arguably one of the best in recent publications. Using complex wavelet representations, the performance of the proposed algorithm is very competitive with that of the state-of-the-art algorithms.

1. INTRODUCTION

Estimating signal from noisy observations is a fundamental task in signal processing. A linear observation model is given by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{x}$ is a column vector of the true signal, $\mathbf{y}$ and $\mathbf{e}$ are vectors of observations and noise, respectively. We assume that $\mathbf{A}$ is a known matrix. When noise is assumed independent and identically distributed (i.i.d.) Gaussian, the maximum likelihood (ML) estimation is a typical least-squares problem [1]. When the distribution of the noise is assumed heavy-tailed, the ML estimation is a robust regression problem [2]. When noise is i.i.d. Gaussian, the maximum a posteriori (MAP) estimation problem is essentially a penalized least-squares problem. For example, the problem is known as a ridge-regression [3] or weight-decay [4] problem when the prior for $\mathbf{x}$ is also i.i.d. Gaussian. When the prior for $\mathbf{x}$ is non Gaussian, it is usually chosen to model the sparseness of the signal [5, 6], to control the complexity of the model [3, 4], or to perform model selection [7, 8]. A typical application that exploits the sparseness of the signal is in wavelet-based image denoising [9–11].

Among the non-Gaussian distributions, a particular family of heavy-tailed distributions is the scale mixture of Gaussian (SMG) distribution [12]. These distribution functions, including power exponential, student-t, and slash distributions [13, 14], have found many successful applications in robust statistical data analysis [14, 15], image processing [11, 16], and machine learning problems [6, 17]. An interesting discussion of robust estimation using the SMG distribution is given in [18].

After we make assumptions about the likelihood and the prior, the MAP estimate of $\mathbf{x}$ can be determined by solving a specific optimization problem. Optimization algorithms

that are based on variational methods [19, 20] and the minorization-maximization principle [21] are powerful alternatives to the EM algorithm. For example, optimization algorithms based on variational methods have been widely used in Bayesian learning that employs non-Gaussian distributions. These algorithms are based on maximizing the variational lower bound which is given either by introducing an auxiliary distribution function [20, 22, 23] or by representing the objective function in its variational form using the conjugate function [19, 24, 25]. Variational methods have also been applied to a number of signal processing problems [17, 23, 26]. The basic idea of the minorization-maximization (MM) algorithm [21] is that, instead of directly maximizing the objective function, another objective function that minorizes the original objective function is iteratively maximized. (The formal definition is presented in Section 2.2.) The MM algorithm has been successfully applied to solve many statistical problems including variable selection [27] and quantile regression [28]. It has also found applications in machine learning research [29]. Both variational methods and the MM algorithm have long been applied to solve many signal processing problems such as image restoration [30–34] and computer tomography [35–38].

In this paper, we develop an iterative algorithm to determine the MAP estimate of $\mathbf{x}$ based on the Gaussian linear observation model given by (1) and the prior $p(\mathbf{x})$ given by an i.i.d. scale mixture of Gaussian distribution. A direct application of this algorithm is image denoising in the wavelet domain. This is performed by letting $\mathbf{A} = \mathbf{I}$ in the observation model and regarding $\mathbf{y}$ and $\mathbf{x}$ as the observed and original wavelet coefficients, respectively. This is a special case of the linear observation model. Since our study is based on the conjugate function lower bound (CFLB) and minorization-maximization, we also study the connection between the two.

This paper is organized as follows. In Section 2, after a brief introduction of the CFLB and MM algorithms, we present a generalized view of both algorithms and two extensions. In particular, we study iterative optimization algorithms for a class of objective functions $F(x)$. We assume that through a suitable mapping of the variable $t = q(x)$, the resulting function $f(t)$ ($f[q(x)] = F(x)$) is convex and differentiable. This type of objective function is studied because the log-prior (the logarithm of the scale mixture of Gaussian distribution) has this property and plays an important role in the algorithmic development. We describe iterative algorithms called the CFLB algorithm and MM algorithm for maximizing the objective function. We then propose a generalized version of both algorithms and show that they are indeed equivalent for the class of objective functions considered in this paper. In Section 3, we develop an iterative algorithm for MAP estimate of the signal under the general model setting of (1). We also discuss the connection between the developed algorithm with an EM algorithm. In Section 4, we modify the algorithm developed in Section 3 for image denoising in the wavelet domain. We study two heavy-tailed priors for wavelet coefficients: student-t and slash distributions which are of interest as they have not yet been widely studied in image denoising. Recognizing that the proposed algorithms can be regarded as generalized Wiener estimation algorithms, we propose two algorithms which exploit the local statistics. One is a noniterative algorithm which has a parameter that accounts for the heavy-tailed characteristics of the signal. The other is an iterative algorithm based on either the student-t or slash distribution. We also discuss the connection of proposed algorithms with algorithms based on empirical Bayes and issues related to using the proposed algorithm in complex wavelet representations. Experimental results show that when using a single wavelet representation the performance of the proposed algorithms is better than that of the bi-shrinkage algorithm [39] which is arguably one of the best in recent publications. Using over-complete wavelet representations, the performance of the proposed algorithm is competitive to that of the state-of-the-art image denoising algorithms [16, 39, 40].

2. ITERATIVE MAXIMIZATION BASED ON THE CONVEXITY OF THE OBJECTIVE FUNCTION

In this section, we present a brief introduction to the CFLB and MM algorithms for determining the local/global maximum of a objective function $F(x)$. We assume that through a suitable mapping $t = q(x)$ we have a convex and differentiable function $f(t)$ such that $F(x) = f[q(x)] = f(t)$. The proposed CFLB and MM algorithms are based on the convexity of $f(t)$. We then present a generalized version of these two algorithms and two extensions. This is followed by two families of objective functions for which the CFLB and MM algorithms are useful tools.

2.1. The conjugate function lower bound 85(CFLB) algorithm [19, 20]

The conjugate function [41] of $f(t)$ is

$$f^*(\lambda) = \arg\max_t [\lambda t - f(t)]. \tag{2}$$

When $f(t)$ is convex and differentiable, the maximizer of $\lambda t - f(t)$ satisfies $\lambda = f'(t)$, where $f'(t) = df(t)/dt$. For a fixed t, $f(t)$ is recovered by

$$f(t) = \arg\max_\lambda [\lambda t - f^*(\lambda)]. \tag{3}$$

Using Fenchel's inequality [41], for any t and λ, the conjugate function lower bound for $f(t)$ is given by

$$f(t) \geq \lambda t - f^*(\lambda) \quad \text{for any } \lambda. \tag{4}$$

We can define a new objective function

$$P(x, \lambda) = \lambda q(x) - f^*(\lambda). \tag{5}$$

Substituting $t = q(x)$ into (3), it is clear that

$$F(x) = \arg\max_{\lambda} P(x,\lambda) \tag{6}$$

and $F(x) \geq P(x,\lambda)$ for any λ.

An iterative algorithm that guarantees a nondecreasing sequence of $F(x)$ is the following. The algorithm, called the conjugate function lower bound (CFLB) algorithm, has two maximization steps. At the kth iteration, we know $x^{(k)}$ and maximize $P(x^{(k)},\lambda)$ to obtain $\lambda^{(k)}$. It is easy to show that

$$\lambda^{(k)} = \arg\max_{\lambda} P(x^{(k)},\lambda) = f'(t^{(k)}), \tag{7}$$

where $t^{(k)} = q(x^{(k)})$. Next, we calculate $x^{(k+1)}$ by maximizing $P(x,\lambda^{(k)})$,

$$x^{(k+1)} = \arg\max_{x} P(x,\lambda^{(k)}), \tag{8}$$

where we assume that there is at least a local maximum for $P(x,\lambda^{(k)})$. Since $\lambda^{(k)}$ is fixed, $f^*(\lambda^{(k)})$ is a constant. We can write

$$P(x,\lambda^{(k)}) = \lambda^{(k)} q(x) + \text{constant}. \tag{9}$$

From the above two maximization steps, we can write

$$P(x^{(k)},\lambda^{(k)}) \leq P(x^{(k+1)},\lambda^{(k)}) \leq P(x^{(k+1)},\lambda^{(k+1)}). \tag{10}$$

According to definition, we have $F(x^{(k)}) = P(x^{(k)},\lambda^{(k)})$ and $F(x^{(k+1)}) = P(x^{(k+1)},\lambda^{(k+1)})$. Thus, $F(x^{(k+1)}) \geq F(x^{(k)})$. Therefore, the CFLB algorithm leads to a nondecreasing sequence of $F(x^{(k)})$.

2.2. The minorization-maximization (MM) algorithm

A function $g(t;t^{(k)})$ with a known parameter $t^{(k)}$ is said to minorize $f(t)$ at the point $t^{(k)}$ provided

$$\begin{aligned} g(t;t^{(k)}) &\leq f(t) \quad \forall t \\ g(t^{(k)};t^{(k)}) &= f(t^{(k)}). \end{aligned} \tag{11}$$

Let $t^{(k+1)}$ be the maximizer of $g(t;t^{(k)})$, such that

$$t^{(k+1)} = \arg\max_{t} g(t;t^{(k)}). \tag{12}$$

From the definition, we have

$$\begin{aligned} g(t^{(k+1)};t^{(k)}) &\geq g(t^{(k)};t^{(k)}) = f(t^{(k)}), \\ f(t^{(k+1)}) &\geq g(t^{(k+1)};t^{(k)}). \end{aligned} \tag{13}$$

Therefore, maximizing $g(t;t^{(k)})$ results in a nondecreasing sequence $f(t^{(k+1)}) \geq f(t^{(k)})$. This algorithm is called a minorization-maximization (MM) algorithm.

For a convex and differentiable function $f(t)$, we have

$$f(t) \geq f(t^{(k)}) + f'(t^{(k)})(t - t^{(k)}). \tag{14}$$

Substituting $t = q(x)$ into (14), we have

$$F(x) \geq F(x^{(k)}) + f'[q(x^{(k)})][q(x) - q(x^{(k)})]. \tag{15}$$

Thus, $F(x)$ is minorized by

$$\begin{aligned} D(x;x^{(k)}) &= F(x^{(k)}) + f'[q(x^{(k)})][q(x) - q(x^{(k)})] \\ &= f'[q(x^{(k)})]q(x) + \text{constant}, \end{aligned} \tag{16}$$

since $F(x^{(k)}) = D(x^{(k)};x^{(k)})$ and $F(x) \geq D(x;x^{(k)})$. Therefore, the MM algorithm that iteratively maximizes $D(x;x^{(k)})$,

$$x^{(k+1)} = \arg\max_{x} D(x;x^{(k)}), \tag{17}$$

leads to a nondecreasing sequence $F(x^{(k+1)}) \geq F(x^{(k)})$. The convergent property of the MM algorithm and techniques to speed up the convergence rate are discussed in [21, 42].

2.3. Generalization, comparison, and extensions

2.3.1. Generalization and comparison

The basic ideas of the CFLB and MM algorithms can be generalized as the following. To find a local/global maximum of $f(x)$, we assume there is a function $d(x,y)$:

$$\begin{aligned} f(x) &\geq d(x,y) \quad \text{for any } x, y, \\ f(x) &= d(x,y_x) \quad \text{for a given } x, \end{aligned} \tag{18}$$

where $y_x = h(x)$ is a suitable function of x. An iterative algorithm can now be developed. In the kth step, we assume $x^{(k)}$ is known and we calculate $y_x^{(k)} = h(x^{(k)})$. Then in the $(k+1)$th step, we determine $x^{(k+1)}$ such that

$$d(x^{(k+1)},y_x^{(k)}) \geq d(x^{(k)},y_x^{(k)}). \tag{19}$$

Since by definition $f(x^{(k+1)}) \geq d(x^{(k+1)},y_x^{(k)})$ and $f(x^{(k)}) = d(x^{(k)},y_x^{(k)})$, we have $f(x^{(k+1)}) \geq f(x^{(k)})$. Therefore, when the two conditions stated in (18) are satisfied, the uphill location $x^{(k+1)}$ for $d(x,y_x^{(k)})$ is also the uphill location for $f(x)$.

The CFLB and MM algorithms are special cases of the above generalized algorithm. There are two special considerations in the CFLB and MM algorithms.

(i) One is operational. $x^{(k+1)}$ is determined by the maximization $x^{(k+1)} = \max_x d(x,y_x^{(k)})$. In light of (19), this maximization step in the CFLB and MM algorithm is sufficient but not necessary.

(ii) The other is structural. For an MM algorithm, $y_x^{(k)} = x^{(k)}$, while for a CFLB algorithm, the function $y_x^{(k)} = h(x^{(k)})$ depends on the definition of the conjugate function lower bound $d(x,y)$.

In addition, we can clearly see that the objective function $P(x,\lambda^{(k)})$ of the CFLB algorithm is exactly the same as the minorization function $D(x;x^{(k)})$ of the MM algorithm.

This is because for a convex and differentiable function its conjugate function lower bound [41] is the same as the minorizing function used in the MM algorithm. Therefore, the CFLB algorithm and the MM algorithm, when they rely on the convexity of $f(t)$, are essentially the same in searching for a local/global maximum of the function $F(x)$. We note that for the MM algorithm there are other tools to construct the minorizing function which is not necessarily the same as that constructed by using the CFLB.

2.3.2. Two extensions

Here, we assume the objective function has at least a local maximum. In the first extension, we consider an objective function of a vector variable $\mathbf{x}$:

$$J_1(\mathbf{x}) = \sum_{n=1}^{N} F_n(\mathbf{x}), \tag{20}$$

where $F_n(\mathbf{x}) = f[q_n(\mathbf{x})]$ and $q_n(\mathbf{x}) = t_n$ is scalar. Assume $f(t)$ is convex. Then, $J_1(\mathbf{x})$ is minorized by $G(\mathbf{x};\mathbf{x}^{(k)}) = \sum_{n=1}^{N} f'(t_n^{(k)})q_n(\mathbf{x})$. The CFLB/MM algorithm for maximizing $J_1(\mathbf{x})$ is given by

$$\mathbf{x}^{(k+1)} = \arg\max_{\mathbf{x}} G(\mathbf{x};\mathbf{x}^{(k)}). \tag{21}$$

In the second extension, we consider an objective function which is the sum of $J_1(\mathbf{x})$ defined in (20) and another objective function $J_0(\mathbf{x})$:

$$J(\mathbf{x}) = J_0(\mathbf{x}) + J_1(\mathbf{x}). \tag{22}$$

Here, we assume that $J(\mathbf{x})$ has at least a local maximum. Since $J_1(\mathbf{x})$ is minorized by $G(\mathbf{x};\mathbf{x}^{(k)})$, $J(\mathbf{x})$ is minorized by $H(\mathbf{x};\mathbf{x}^{(k)}) = J_0(\mathbf{x}) + G(\mathbf{x};\mathbf{x}^{(k)})$. The CFLB/MM algorithm for maximizing $J(\mathbf{x})$ is given by

$$\mathbf{x}^{(k+1)} = \arg\max_{\mathbf{x}} H(\mathbf{x};\mathbf{x}^{(k)}). \tag{23}$$

2.4. An example

In this section, we show that the logarithm of the scale mixture of Gaussian distribution function [12] has the desired convex property after suitable mapping of variables. A family of heavy-tailed distributions for a zero-mean scalar random variable x is defined as a scale mixtures of Gaussian:

$$p(x \mid \sigma^2, \nu) = \int_0^\infty \mathcal{N}(x \mid \sigma^2, u)\, p(u \mid \nu) du, \tag{24}$$

where $\mathcal{N}(x \mid \sigma^2, u) = (\sqrt{u}/\sqrt{2\pi}\sigma)e^{-(u/2\sigma^2)x^2}$ and $p(u \mid \nu)$ is the prior distribution of $u\,(0 \le u < \infty)$. (We adopt the following notations in this paper. The conditional distribution function is denoted by $p(x \mid y)$, while a function parameterized by a parameter is denoted by $f(x; y)$. A vector $\mathbf{x}$ is written in bold-face font, while its nth elements is denoted by x_n.) Here, we have included two parameters σ^2 and ν to account for certain distributions such as the student-t distribution which has a scaling parameter and a parameter for the degree of freedom. Different settings for $p(u \mid \nu)$ result in a family of heavy-tailed distributions [11, 12, 18]. For example, when $p(u \mid \nu)$ is a gamma distribution with both parameters set to $\nu/2$, the resulting SMG is the student-t distribution. The Gaussian distribution is a special case where u is not a random variable but is a constant $u = 1$. The definitions of three heavy-tailed distributions: power exponential, student-t, and slash are shown in Table 1. More examples can be found in [11]. We note that the Laplacian ($\nu = 1$) and Gaussian distribution ($\nu = 2$) are two special cases of the power exponential distribution. In addition, the power exponential distribution function can be represented as the SMG when $0 < \nu \le 2$. (This is a subset of the power exponential distribution function that can be represented as SMG.)

Table 1: Three heavy-tailed distributions are presented, where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ and $\Gamma(a,b) = \int_0^b t^{a-1}e^{-t}dt$ are the gamma function and incomplete gamma function, respectively. We assume ν is a fixed parameter.

Power exponential	$\frac{1}{2\Gamma(1+1/\nu)\sigma}\exp\left[-(x^2/\sigma^2)^{\nu/2}\right], 0 < \nu \le 2$
Student-t	$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma}(1 + x^2/\nu\sigma^2)^{-(\nu+1)/2}$
Slash	$\frac{\nu}{\sqrt{2\pi}\sigma}(x^2/2\sigma^2)^{-(\nu+1/2)}\Gamma(\nu+1/2, x^2/2\sigma^2)$

We will use $s = \sigma^2$ to simplify notations in the following discussion. We study the logarithm of the scale mixture of Gaussian distribution function

$$\log\int_0^\infty \mathcal{N}(x \mid s, u)p(u \mid \nu)du = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log s + F(x), \tag{25}$$

where

$$F(x) = \log\int_0^\infty \sqrt{u}p(u \mid \nu)\exp\left[-\frac{ux^2}{(2s)}\right]du. \tag{26}$$

Changing the variable $t = q(x) = x^2/(2s)$, we have a new function $f(t)$ and its first derivative as follows:

$$f(t) = \log\int_0^\infty \sqrt{u}p(u \mid \nu)\exp[-ut]du, \tag{27}$$

$$f'(t) = -\frac{\int_0^\infty u^{3/2}p(u \mid \nu)\exp[-ut]du}{\int_0^\infty \sqrt{u}p(u \mid \nu)\exp[-ut]du}. \tag{28}$$

We can verify the convexity of $f(t)$ by recognizing that it is the logarithm of the Laplace transform of $\sqrt{u}p(u \mid \nu)$. The Laplace transform of $\sqrt{u}p(u \mid \nu)$ is log-convex [41]. Therefore, the proposed CFLB/MM algorithm can be used to maximize an objective function which involves log-SMG distribution functions. The function $f(t)$ and its first derivative for the three distributions are listed in Table 2.

3. MAP ESTIMATION UNDER LINEAR GAUSSIAN OBSERVATION MODEL

In this section, we present the development of a CFLB/MM algorithm for solving a general MAP estimation problem

Table 2: The function $f(t)$ and its first derivative for the three heavy-tailed distributions. For power exponential, the parameter ν must be set as $0 < \nu \leq 2$ such that $f(t)$ is convex. We note that $f(t)$ and $f'(t)$ can be directly calculated from the distribution function. The knowledge of the exact form of $p(u \mid \nu)$ is not required.

	Power exponential	Student-t	Slash
$f(t)$	$-(2t)^{\nu/2}$	$-\frac{\nu+1}{2}\log(\nu+2t)$	$-\left(\nu+\frac{1}{2}\right)\log t+\log\Gamma\left(\nu+\frac{1}{2},t\right)$
$f'(t)$	$-\nu(2t)^{(\nu-2)/2}$	$-\frac{\nu+1}{\nu+2t}$	$-\frac{\Gamma(\nu+3/2,t)}{t\Gamma(\nu+1/2,t)}$

of linear Gaussian model. We then discuss the connection between the developed algorithm with an expectation maximization (EM) algorithm.

3.1. Development of the iterative algorithm

Given a linear observation model in (1), we want to determine a MAP estimate of the parameter vector $\mathbf{x}$ based on the following model assumptions. Elements of $\mathbf{e}$ are i.i.d. Gaussian with known variance σ_e^2. Elements of $\mathbf{x}$ are i.i.d. scale mixture of Gaussian with unknown scaling parameter s. The degree of freedom ν is assumed to be a free parameter that can be tuned. (A full Bayesian estimate for ν is generally very computationally complicated [15, 43] and is beyond the scope of this paper.) More specifically, the log-posterior is given by

$$\begin{aligned} J(\mathbf{x},s) &= \log p(\mathbf{x},s \mid \mathbf{y}) \\ &= \log p(\mathbf{y} \mid \mathbf{x}) + \log p(\mathbf{x} \mid s) + \log p(s) + \text{constant}, \end{aligned} \tag{29}$$

where

$$\begin{aligned} \log p(\mathbf{y} \mid \mathbf{x}) &= -\frac{N}{2}\log\left(2\pi\sigma_e^2\right) - \frac{1}{2\sigma_e^2}\mathbf{e}^T\mathbf{e}, \\ \log p(\mathbf{x} \mid s) &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log s + \sum_{n=1}^{N} F_n(\mathbf{x},s), \\ F_n(\mathbf{x},s) &= \log\int_0^\infty \sqrt{u_n}\,p(u_n \mid \nu)\exp\left[-\frac{u_n x_n^2}{(2s)}\right]du_n. \end{aligned} \tag{30}$$

Changing variable $t_n = q_n(\mathbf{x},s) = x_n^2/(2s)$, we have $F_n(\mathbf{x},s) = f[t_n]$, where the function $f(t)$ is convex and is given by (27). We can rewrite (29) as

$$J(\mathbf{x},s) = J_0(\mathbf{x},s) + J_1(\mathbf{x},s), \tag{31}$$

where

$$\begin{aligned} J_0(\mathbf{x},s) &= \log p(\mathbf{y} \mid \mathbf{x}) + \log p(s) - \frac{N}{2}\log s, \\ J_1(\mathbf{x},s) &= \sum_{n=1}^{N} F_n(\mathbf{x},s) = \sum_{n=1}^{N} f(t_n). \end{aligned} \tag{32}$$

Since $J_1(\mathbf{x},s)$ is minorized by $\sum_{n=1}^{N} f'(t_n^{(k)})(x_n^2/2s)$, $J(\mathbf{x},s)$ is minorized by the following objective function:

$$\begin{aligned} &H\left(\mathbf{x},s;\mathbf{x}^{(k)},s^{(k)}\right) \\ &\quad= J_0(\mathbf{x},s) + \sum_{n=1}^{N} f'\left(t_n^{(k)}\right)\frac{x_n^2}{2s} \\ &\quad= -\frac{1}{2\sigma_e^2}\mathbf{e}^T\mathbf{e} + \log p(s) - \frac{N}{2}\log s - \frac{1}{2s}\mathbf{x}^T\mathbf{W}^{(k)}\mathbf{x}, \end{aligned} \tag{33}$$

where $\mathbf{W}^{(k)} = \text{diag}[-f'(t_n^{(k)})]$ is a diagonal matrix. The update for $\mathbf{x}$ is then obtained by maximizing $H(\mathbf{x},s;\mathbf{x}^{(k)},s^{(k)})$,

$$\mathbf{x}^{(k+1)} = \left(\mathbf{A}^T\mathbf{A} + \frac{\sigma_e^2}{s^{(k)}}\mathbf{W}^{(k)}\right)^{-1}\mathbf{A}^T\mathbf{y}. \tag{34}$$

Here, we need to make a further assumption that the matrix $\mathbf{A}$ is properly defined such that the matrix inversion in the above equation can be carried out for each iteration. The next step is to determine the update for the scaling parameter s. To simplify presentation, we assume $p(s)$ is a uniform distribution in this section. Other priors for $p(s)$ are considered in Section 4.1. The update of the scaling parameter is given by

$$s^{(k+1)} = \frac{1}{N}\left[\mathbf{x}^{(k+1)}\right]^T\mathbf{W}^{(k)}\mathbf{x}^{(k+1)}. \tag{35}$$

3.2. Equivalence with the EM algorithm

In [44], we develop an EM algorithm for the MAP estimation problem. In this section, we present the details of the EM algorithm and show that it is equivalent to the CFLB/MM algorithm. In developing the EM algorithm, we regard the parameters $\gamma = \{u_n\}$ as the missing data. The signal $\mathbf{x}$ and the scaling factor s, denoted $\phi = \{\mathbf{x},s\}$, are the data to be estimated. We then determine the Q-function [45]

$$Q(\phi;\phi^{(k)}) = \int_0^\infty \log p(\phi,\gamma \mid \mathbf{y})\,p(\gamma \mid \phi^{(k)},\mathbf{y})\,d\gamma. \tag{36}$$

We now give details of calculating the Q-function and the E-step. Using Bayes' rule, we can write

$$p(\phi,\gamma \mid \mathbf{y}) \propto p(\mathbf{y} \mid \phi,\gamma)\,p(\phi \mid \gamma)\,p(\gamma), \tag{37}$$

where

$$p(\mathbf{y} \mid \phi, \gamma) = \left(\sqrt{2\pi}\sigma_e\right)^{-N} \exp\left(-\frac{\sum_{n=1}^{N} e_n^2}{2\sigma_e^2}\right),$$
$$p(\phi \mid \gamma) = \frac{\prod_{n=1}^{N}\sqrt{u_n}}{\left(\sqrt{2\pi s}\right)^N} \exp\left(-\frac{\sum_{n=1}^{N} u_n x_n^2}{2s}\right) p(s), \quad (38)$$
$$p(\gamma) = \prod_{n=1}^{N} p(u_n).$$

Therefore, ignoring constants and unrelated terms, we have the following results:

$$Q(\phi; \phi^{(k)}) = \log p(\mathbf{y} \mid \mathbf{x}) + \log p(s) - \frac{N}{2}\log s - \frac{1}{2s}\sum u_n^{(k)} x_n^2, \quad (39)$$

where $u_n^{(k)}$ is the conditional mean

$$u_n^{(k)} = E[u_n \mid \phi^{(k)}, \mathbf{y}]. \quad (40)$$

Equation (40) states the calculation required for the E-step. In the M-step, we maximize the Q-function to determine $\phi^{(k+1)}$.

The E-step is calculated as the following:

$$u_n^{(k)} = \int_0^\infty u_n p(u_n \mid x_n^{(k)}, s^{(k)}) du_n, \quad (41)$$

where

$$p(u_n \mid x_n^{(k)}, s^{(k)}) = \frac{p(x_n^{(k)} \mid s^{(k)}, u_n) p(u_n \mid \nu)}{\int_0^\infty p(x_n^{(k)} \mid s^{(k)}, u_n) p(u_n \mid \nu) du_n}. \quad (42)$$

Since $p(x_n^{(k)} \mid s^{(k)}, u_n)$ is Gaussian

$$p(x_n^{(k)} \mid s^{(k)}, u_n) = \frac{\sqrt{u_n}}{\sqrt{2\pi s^{(k)}}} \exp\left[-\frac{u_n (x_n^{(k)})^2}{(2s^{(k)})}\right], \quad (43)$$

we have

$$p(u_n \mid x_n^{(k)}, s^{(k)}) = \frac{\sqrt{u_n} p(u_n \mid \nu) \exp(-u_n t_n^{(k)})}{\int_0^\infty \sqrt{u_n} p(u_n \mid \nu) \exp(-u_n t_n^{(k)}) du_n}, \quad (44)$$

where we have used the substitution $t_n^{(k)} = (x_n^{(k)})^2/(2s^{(k)})$. Substituting (44) into (41) and comparing with (28), we can see that $f'(t^{(k)}) = -u_n^{(k)}$. Comparing objective function of the CFLB/MM algorithm (33) with the Q-function of the EM algorithm (39), we can see the two algorithms are equivalent.

4. APPLICATIONS IN IMAGE DENOISING

4.1. Iterative denoising algorithms

We now use image denoising in wavelet transform domain as an example to demonstrate the application of the proposed algorithm. (Part of this section was presented in [34, 46].) In the wavelet domain, we have the following observation model:

$$\mathbf{y} = \mathbf{x} + \mathbf{e}, \quad (45)$$

where $\mathbf{y}$ and $\mathbf{x}$ are observed and original wavelet coefficients of the signal, respectively. $\mathbf{e}$ is the additive Gaussian noise. Therefore, the denoising problem is a special case of the MAP estimation problem (considered in Section 3), where $\mathbf{A} = \mathbf{I}$ is an identity matrix. From (34), we can easily derive the update for the signal

$$x_n^{(k+1)} = \frac{s^{(k)}}{s^{(k)} + u_n^{(k)}\sigma_e^2} y_n, \quad (46)$$

where we have used $u_n^{(k)} = -f'(t_n^{(k)})$ to simplify notation. (At the end of Section 3.2, we have commented on the relationship between $u_n^{(k)}$ and $f'(t_n^{(k)})$.) The update of the scaling parameter depends on its prior distribution. In this section, we consider three priors: a conjugate prior given by the inverse-chi-square (Inv-χ^2) distribution, Jeffreys' prior ($p(s) \propto 1/s$), and the uniform prior [45]. The Inv-χ^2 distribution is given by

$$p(s \mid \eta) = \frac{2^{-\eta/2}}{\Gamma(\eta/2)} s^{-(\eta/2+1)} \exp\left[-\frac{1}{(2s)}\right], \quad s > 0,\ \eta > 0, \quad (47)$$

where η is the degree of freedom. For $\eta > 2$, the mean of s is given by $E[s] = 1/(\eta - 2)$. Therefore, if we have prior knowledge about the mean of s, say s_0, then $\eta = 2 + 1/s_0$. With these considerations, we can determine the update for the scaling parameter using the Inv-χ^2 prior, Jeffreys' prior and the uniform prior as the following:

$$s^{(k+1)} = \frac{\sum_{n=1}^{N} u_n^{(k)} (x_n^{(k)})^2 + 1}{N + \eta + 2},$$
$$s^{(k+1)} = \frac{1}{N+2}\sum_{n=1}^{N} u_n^{(k)} (x_n^{(k)})^2, \quad (48)$$
$$s^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} u_n^{(k)} (x_n^{(k)})^2.$$

4.2. Generalized wiener estimation

We recall that for the observation model given by (45), when the signal is modeled i.i.d. Gaussian with zero-mean and known variance σ^2, the MAP estimate of x_n is a Wiener estimate given by

$$x_n = \frac{\sigma^2}{\sigma^2 + \sigma_e^2} y_n. \quad (49)$$

To link the proposed iterative algorithm with the Wiener estimation, we compare (46) and (49). It is easy to see that (49) is a special case of (46) where u_n is a constant, that is, $u_n^{(k)} = 1$. We regard the proposed algorithm as a generalized

Wiener estimate, because (a) the variable u_n in (46) is a scaling factor that accounts for the heavy-tailed characteristic of the distribution, and (b) it is an iterative algorithm.

To gain further insight into the proposed algorithm, we study the student-t distribution with the degree of freedom $\nu > 2$. The relationship between the variance of the signal σ_s^2 and the scaling factor s is given by

$$\sigma_s^2 = \frac{\nu}{\nu - 2} s. \tag{50}$$

Thus, once the estimated scale parameter $s^{(k)}$ at the kth iteration is known, the estimated signal variance $(\sigma_s^2)^{(k)}$ is also known. Using this relationship, we can rewrite (46) as

$$x_n^{(k+1)} = \frac{(\nu - 2)(\sigma_s^2)^{(k)} + (x_n^{(k)})^2}{(\sigma_s^2)^{(k)} + (x_n^{(k)})^2 + (\nu + 1)\sigma_e^2} y_n, \tag{51}$$

where we have used $u_n^{(k)} = -f'(t_n^{(k)}) = (\nu + 1)/(\nu + 2t_n^{(k)})$ and $t_n^{(k)} = [x_n^{(k)}]^2/2s$. We can further rewrite (51) as

$$x_n^{(k+1)} = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_e^2} y_n, \tag{52}$$

where

$$\sigma_L^2 = (\sigma_s^2)^{(k)} + \frac{1}{\nu + 1}\left[(x_n^{(k)})^2 - 3(\sigma_s^2)^{(k)}\right]. \tag{53}$$

Comparing (49) and (52), we can see that the latter can be regarded as a generalized Wiener estimate of the signal, where a localized signal variance σ_L^2 is estimated by taking a weighted average of the signal variance σ_s^2 and the local signal energy. It can be easily seen that when $\nu \to \infty$, the student-t distribution approaches the Gaussian distribution and $\sigma_L^2 = (\sigma_s^2)^{(k)}$. In this case, (52) is a generalized form of (49) in that it represents an iterative algorithm for estimating signal under unknown signal variance.

4.3. Two image denoising algorithms

Direct application of the proposed algorithm for image denoising does not necessarily lead to satisfactory results. This is because in developing the algorithm we have ignored that image signals are generally nonstationary. Since the proposed algorithms can be regarded as generalized Wiener estimates that use localized information, they are modified in the following two ways for image denoising.

4.3.1. A noniterative generalized Wiener estimation algorithm

Motivated by developing a low-complexity algorithm, we consider a noniterative algorithm given by

$$x_n = \frac{\sigma_n^2}{\sigma_n^2 + \alpha\sigma_e^2} y_n, \tag{54}$$

where σ_n^2 is a localized estimate of the signal energy at the nth location and α is constant to be determined for a particular class of signals. When $\alpha = 1$, this algorithm is a Wiener estimate using local statistics. The heavy-tail distribution of the signal is accounted for by setting $\alpha \neq 1$. The performance of this algorithm also depends on the estimation of noise variance σ_e^2 and the local signal variance σ_n^2. A robust estimation of the variance [10] of the noise is given by

$$\sigma_e = \frac{\text{median}(|\mathbf{y}|)}{0.6745}. \tag{55}$$

A simple method to estimate the signal variance is the following:

$$\sigma_n^2 = \begin{cases} S_n - \sigma_e^2, & S_n > \sigma_e^2, \\ 0, & \text{otherwise}, \end{cases} \tag{56}$$

where $S_n = (1/(2M+1))\sum_{k=-M}^{M} y_{n-k}^2$. The underlying principle for this estimation is that the signal is uncorrelated with noise. With the above results, we can see that the proposed algorithm (in (54)) is actually a combination of shrinkage and hard-thresholding. The shrinkage part is a generalized adaptive Wiener filter with the parameter α accounting for the heavy-tailed characteristics of the signal, while the hard-thresholding part, which is well established [10], plays an essential role in favouring a sparse solution.

It should be noted that (54) is not a direct result of an optimization problem. It is, however, a low-complexity approximation of the iterative algorithm given by (46). Indeed, comparing these two equations, we can see that $\mu_n^{(k)}$ and $s_n^{(k)}$ in (46) are replaced by α and σ_n^2, respectively. As such, we can regard (46) as a one-step implementation of the iterative algorithm.

4.3.2. An iterative generalized Wiener estimation algorithm using local statistics

This algorithm is motivated by using the local statistics discussed in Section 4.2. We note that in developing (46) we have assumed a global scaling parameter s for the whole image. This assumption is useful to simplify discussion. However, it is not necessarily a valid one for wavelet coefficients of an image. Therefore, we propose to replace the global scaling parameter s with a localized scaling parameter s_n which is estimated by

$$s_n^{(k+1)} = \frac{1}{2M+1} \sum_{m=-M}^{M} u_{n-m}^{(k)} \left[x_{n-m}^{(k+1)}\right]^2. \tag{57}$$

From Table 2, we can see that $u_n^{(k)}$ is a function of $[x_n^{(k)}]^2/s^{(k)}$ for the student-t and slash distribution. We replace it with $z_n^{(k)}/s_n^{(k)}$, where

$$z_n^{(k)} = \frac{1}{2M+1} \sum_{m=-M}^{M} \left[x_{n-m}^{(k)}\right]^2. \tag{58}$$

The estimate of the signal is then given by

$$x_n^{(k+1)} = \frac{s_n^{(k)}}{s_n^{(k)} + u_n^{(k)}\sigma_e^2} y_n. \tag{59}$$

Table 3: PSNR (dB) results using two noisy images with different levels of additive noise. GWE1 and GWE2 are the GWE algorithms with $\alpha = 1$ and $\alpha = \sqrt{2}$, respectively.

Barbara					
σ_e	GWE1	GWE2	IGWE-T	IGWE-S	Bishrink [47]
10	33.10	32.84	33.05	33.03	32.25
15	30.69	30.62	30.78	30.76	29.97
20	28.97	29.07	29.22	29.20	28.36
25	27.65	27.89	28.03	28.01	27.16
30	26.58	26.98	27.11	27.09	26.28
35	25.71	26.19	26.32	26.31	—
40	24.92	25.56	25.67	25.67	—
Lena					
σ_e	GWE1	GWE2	IGWE-T	IGWE-S	Bishrink [47]
10	34.56	34.63	34.81	34.79	34.36
15	32.37	32.81	32.92	32.89	32.51
20	30.72	31.48	31.54	31.51	31.19
25	29.50	30.45	30.48	30.44	30.15
30	28.37	29.61	29.60	29.56	29.41
35	27.43	28.88	28.84	28.80	—
40	26.62	28.28	28.20	28.16	—

Comparing (59) with (46), we can see that we have used a local estimate of the scaling parameter to replace the global scaling parameter.

4.3.3. Experimental results

The noniterative and the iterative algorithms will be referred to as generalized Wiener estimate (GWE) and iterative generalized Wiener estimate (IGWE), respectively. For the GWE algorithm, extensive experiments using different images have shown that setting $\alpha = \sqrt{2}$ has led to good results in terms of the peak-signal-to-noise ratio (PSNR) of the denoised image. For the IGWE algorithm, since the power exponential and a number of SMG distributions have been studied [11, 16, 48], we focus on the student-t and slash distributions which have not been widely applied to denoising problem. We use IGWE-T and IGWE-S to indicate the student-t and slash distributions being used, respectively. Experimental results show that good results are obtained for 3 to 4 iterations for the IGWE-T ($\nu = 3$) and IGWE-S ($\nu = 15$) algorithms.

We first test image denoising using a single wavelet representation. In all experiments, an image is decomposed into 6 levels using the `sym12` wavelet. Each subband of the signal is then denoised independently. We have performed simulations using the Barbara and Lena images. The experimental results for each noise level setting are obtained by taking the average of the PSNR of 100 runs of the program. In each run of the program, pseudo-Gaussian noise generator is reset to a different state and noise is added to the image.

We can see from the experimental results shown in Table 3 that for the Barbara image the iterative algorithms perform better than the noniterative algorithms. For the Lena image, their performance is about the same. We can also see that using the GWE algorithm, the PSNR associated with the setting $\alpha = \sqrt{2}$ is generally higher than that with the setting $\alpha = 1$. The difference in PSNR is significant for images with high-noise levels. We also note that slight improvement in PSNR can be achieved by varying the value of α between 1.2 to 1.5 according to the estimated noise variance.

We compare the performance of the proposed algorithms with that of the bi-shrinkage [39] which is arguably one of the best in recent publications. We can see that the performance of proposed algorithms (GWE2, IGWE-T, and IGWE-S) is consistently better than that of the bi-shrinkage algorithm.

Next, we test the proposed algorithm using the complex wavelet representation [39]. In our experiments, we use the proposed algorithms (IGWE-T and IGWE-S) to process each individual image subband of the complex wavelet representation. We use exactly the same complex wavelet transform as that used in [39]. For IGWE-T and IGWE-S, the number of iterations is 3 and the degrees of freedom are set $\nu = 3$ (IGWE-T) and $\nu = 15$ (IGWE-S). Results are shown in Table 4.

We can see that the performances of the two proposed iterative algorithms are almost the same. When we compare the results of the proposed algorithms in Table 4 and with respective results in Table 3, we can see that using the complex wavelet representation has led to substantially improved results. Next, we compare results from three image denoising algorithms which use different over complete wavelet representations and different statistical models [16, 39, 40]. We can see from Tables 4 and 5 that the performance of the proposed algorithms are comparable with that of the three

Table 4: A comparison of denoising results based on the peak-signal-to-noise ratio (dB). Five test images are used under different noise levels. It should be noted that results due to references [16, 39, 40] are calculated using the available software from the authors. These results may be slightly different from those presented in the original paper.

	IGWE-S	IGWE-T	[40]	[16]	[39]
Lena					
$\sigma_e = 10$	35.30	35.33	35.0	35.60	35.34
$\sigma_e = 15$	33.47	33.51	33.12	33.90	33.67
$\sigma_e = 20$	32.13	32.18	31.76	32.67	32.40
$\sigma_e = 25$	31.06	31.13	30.69	31.69	31.40
$\sigma_e = 30$	30.18	30.25	29.80	30.87	30.54
$\sigma_e = 40$	28.78	28.83	28.40	29.61	29.23
$\sigma_e = 50$	27.69	27.74	27.30	28.62	28.21
Barbara					
$\sigma_e = 10$	33.69	33.70	33.45	34.03	33.67
$\sigma_e = 15$	31.52	31.54	31.20	31.86	31.47
$\sigma_e = 20$	29.97	29.99	29.64	30.32	29.93
$\sigma_e = 25$	28.78	28.80	28.44	29.13	28.74
$\sigma_e = 30$	27.83	27.85	27.48	28.14	27.80
$\sigma_e = 40$	26.36	26.37	26.0	26.62	26.39
$\sigma_e = 50$	25.26	25.27	24.90	25.46	25.32
Boat					
$\sigma_e = 10$	33.13	33.17	33.09	33.58	33.23
$\sigma_e = 15$	31.21	31.26	31.05	31.70	31.37
$\sigma_e = 20$	29.84	29.88	29.65	30.38	30.02
$\sigma_e = 25$	28.78	28.83	28.58	29.36	29.0
$\sigma_e = 30$	27.94	27.98	27.72	28.55	28.18
$\sigma_e = 40$	26.65	26.69	26.40	27.30	26.92
$\sigma_e = 50$	25.69	25.72	25.42	26.37	26.0
Peppers					
$\sigma_e = 10$	34.95	34.99	34.75	35.36	34.98
$\sigma_e = 15$	33.36	33.42	33.07	33.91	33.49
$\sigma_e = 20$	32.13	32.19	31.84	32.82	32.34
$\sigma_e = 25$	31.14	31.21	30.87	31.93	31.43
$\sigma_e = 30$	30.30	30.34	30.04	31.18	30.66
$\sigma_e = 40$	28.93	29.0	28.74	29.94	29.41
$\sigma_e = 50$	27.85	27.91	27.30	28.99	28.41
Mandrill					
$\sigma_e = 10$	27.98	27.96	30.64	30.78	28.54
$\sigma_e = 15$	26.61	26.60	28.75	28.45	27.02
$\sigma_e = 20$	25.60	25.59	26.74	26.91	25.89
$\sigma_e = 25$	24.83	24.82	25.63	25.80	25.03
$\sigma_e = 30$	24.22	24.21	24.76	24.93	24.34
$\sigma_e = 40$	23.31	23.30	23.52	23.70	23.34
$\sigma_e = 50$	22.66	22.65	22.65	22.87	22.63

Table 5: A comparison of denoising results based on the mean absolute error. Five test images are used under different noise levels. It should be noted that results due to references [16, 39, 40] are calculated using the available software from the authors.

	IGWE-S	IGWE-T	[40]	[16]	[39]
Lena					
$\sigma_e = 10$	3.25	3.24	3.40	3.17	3.23
$\sigma_e = 15$	3.95	3.93	4.14	3.78	3.95
$\sigma_e = 20$	4.58	4.55	4.80	4.31	4.52
$\sigma_e = 25$	5.17	5.13	5.40	4.78	5.03
$\sigma_e = 30$	5.72	5.68	5.96	5.21	5.51
$\sigma_e = 40$	6.72	6.67	6.99	5.98	6.35
$\sigma_e = 50$	7.65	7.58	7.95	6.68	7.12
Barbara					
$\sigma_e = 10$	3.96	3.94	4.09	3.82	3.98
$\sigma_e = 15$	5.02	5.0	5.23	4.84	5.06
$\sigma_e = 20$	5.96	5.94	6.21	5.72	5.99
$\sigma_e = 25$	6.82	6.80	7.09	6.54	6.83
$\sigma_e = 30$	7.60	7.58	7.90	7.28	7.58
$\sigma_e = 40$	9.02	9.0	9.36	8.65	8.91
$\sigma_e = 50$	10.26	10.23	10.63	9.83	10.05
Boat					
$\sigma_e = 10$	4.31	4.29	4.37	4.13	4.30
$\sigma_e = 15$	5.28	5.26	5.43	5.03	5.24
$\sigma_e = 20$	6.11	6.08	6.30	5.77	6.04
$\sigma_e = 25$	6.83	6.80	7.05	6.42	6.73
$\sigma_e = 30$	7.49	7.45	7.73	6.99	7.33
$\sigma_e = 40$	8.62	8.58	8.91	7.97	8.36
$\sigma_e = 50$	9.60	9.56	9.92	8.79	9.22
Peppers					
$\sigma_e = 10$	3.43	3.41	3.55	3.31	3.46
$\sigma_e = 15$	4.06	4.03	4.24	3.85	4.05
$\sigma_e = 20$	4.63	4.60	4.83	4.32	4.59
$\sigma_e = 25$	5.17	5.12	5.36	4.75	5.06
$\sigma_e = 30$	5.68	5.63	5.87	5.14	5.50
$\sigma_e = 40$	6.63	6.57	6.79	5.87	6.31
$\sigma_e = 50$	7.52	7.45	7.56	6.51	7.05
Mandrill					
$\sigma_e = 10$	7.79	7.81	5.83	5.73	7.31
$\sigma_e = 15$	9.05	9.06	7.56	7.38	8.63
$\sigma_e = 20$	10.10	10.11	8.95	8.72	9.76
$\sigma_e = 25$	10.99	11.0	10.12	9.85	10.73
$\sigma_e = 30$	11.75	11.77	11.14	10.83	11.57
$\sigma_e = 40$	13.02	13.03	12.80	12.42	12.93
$\sigma_e = 50$	14.02	14.03	14.12	13.66	14.01

state-of-the-art algorithms in terms of the peak-signal-to-noise ratio and mean absolute error.

The mandrill (also known as baboon) image is quite different from the other four test images in that it contains a lot of fine details. We notice that when the noise level is low ($\sigma_e < 25$) the performances of the two iterative algorithms are not as good as those of three published algorithms. This may be because the window size (7×7) used in the calculation of local signal variance does not match the characteristics of the image. Another reason could be that the prior with a fixed setting of parameter ν does not model the signal well. Therefore, the proposed algorithm could be improved by

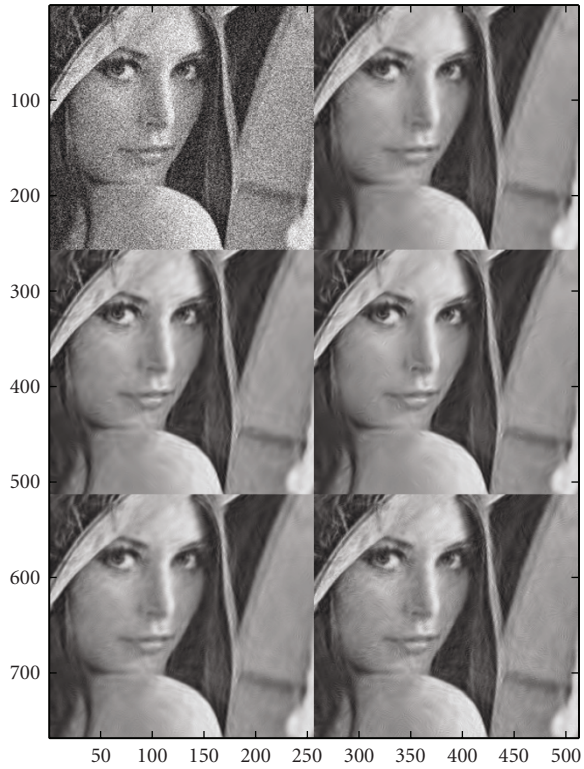

Figure 1: Image denoising results using the Lena image added with random noise ($\sigma_e = 25$). Images shown from top to bottom, left to right are the noisy results of algorithms in [16, 39, 40], the proposed IGWE-T and GWE2 algorithms. Typical PSNR values for these images are listed in Table 4.

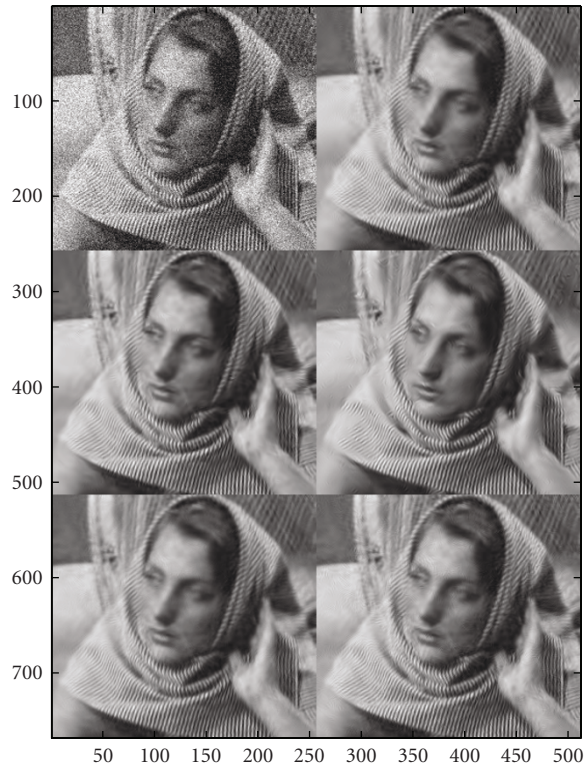

Figure 2: Image denoising results using the Barbara image added with random noise ($\sigma_e = 25$). Images shown from top to bottom, left to right are the noisy, results of algorithms in [16, 39, 40], the proposed IGWE-T and GWE2 algorithms. Typical PSNR values for these images are listed in Table 4.

introducing an adaptive estimation of the window size and the parameter ν. However, this may significantly increase the computational cost.

In Figures 1 and 2, we compare the results of denoised images using algorithms in [16, 39, 40] and the proposed IGWE-T and GWE2 algorithms. (Professor Xin Li kindly provided us with his source code. Matlab codes for the algorithms in [16, 39] are available from the following addresses: http://decsai.ugr.es/~javier/denoise/software/ and http:// taco.poly.edu/WaveletSoftware/, respectively. Default settings for algorithms in [39, 40] are used and suggested settings to reproduce results in [16] are also used.) Again, we can see that the denoised images are quite similar. We note that computation time of these algorithms are quite different in our simulations using a PC with a Pentium 4 3 GHz processor. While the running time for algorithm in [16] is more than 75 seconds those for the proposed GWE2 and IGWET algorithms are about 2.5 and 3.7 seconds, respectively. The running time for the algorithm in [39] is about 3.6 seconds and the running time for the algorithm in [40] is about 7.2 seconds.

4.4. Discussion

In [49], an empirical Bayes (EB) approach is proposed to develop low-complexity image denoising algorithms in which parameters of the prior are estimated from the data. These estimated parameters are then "plugged" into the posterior. The proposed iterative algorithms using local statistics can be regarded as a generalization of the idea of [49] in that the scaling parameter is treated as a random variable and is jointly estimated with the signal. More specifically, the difference is in the way the problem is formulated. For the denoising problem considered in this paper, if we used an EB approach, we would first determine an estimate (e.g., a MAP estimate) of the scale parameter s from the marginal distribution $\hat{s} = \arg\max_s p(s \mid \mathbf{y})$, where $p(s \mid \mathbf{y}) = \int p(\mathbf{x}, s \mid \mathbf{y}) d\mathbf{x}$. We would then determine an estimate (e.g., a MAP estimate) of the signal by assuming a known scale parameter $\hat{s}$, that is, $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}, \hat{s})$. The approach used in this paper, however, is different in that we determine a MAP estimate from the joint posterior $p(\mathbf{x}, s \mid \mathbf{y})$ by using the proposed iterative algorithm.

Another interesting question is as follows. In the observation model (see (45)), it is assumed that the wavelet transform is orthogonal. However, the complex wavelet transform is redundant and is usually nonorthogonal. Can we still apply the denoising method developed for signals in the orthogonal wavelet transform domain to signals in the complex wavelet transform domain? This question is partly answered in a recent paper [50] by Elad. Elad showed that for signal denoising using redundant representations an iterative

algorithm such as the basis pursuit [51] is usually employed. Elad further showed that applying a shrinkage function (usually developed for orthogonal wavelet representations) to redundant wavelet representations is justified in that this can be regarded as the first iteration step of the basis pursuit algorithm.

In addition, the number of iterations deserves further study. As the proposed iterative algorithm is essentially an EM algorithm, it may converge to a local minimum. On the other hand, since we use a local neighborhood to update the scale parameter s, we effectively make a further assumption that the scale parameter also follows an i.i.d. distribution locally. This assumption may fit the data well in the first few iterations. But this may not be the case after a few iterations when the signal is less noisy. This is perhaps an intuitive explanation of the observation that the performance (measured by the PSNR and mean absolute errors) of the proposed iterative algorithm improves in the first 3 to 4 iterations, drops slightly, and converges to a suboptimal estimate.

The proposed algorithms are not optimal in removing non-Gaussian noise (e.g., impulsive noise). This is because we have taken a model-based approach in solving a MAP estimating problem involving a Gaussian linear observation model which has been used in many recent publications (see, e.g., [11] and reference therein). As such, the solution is only optimal for Gaussian noise. From a model-based point of view, to deal with non-Gaussian noise, we need to make proper assumption about the noise distribution function and solve the MAP problem. Such work is beyond the scope of this paper.

5. CONCLUSIONS

In this paper, we have studied CFLB/MM algorithms for a special class of objective functions that are convex through a suitable mapping of variable. We proposed a generalized version of the CFLB/MM algorithm and show that the CFLB and MM algorithms are equivalent for this class of objective functions. We develop a CFLB/MM algorithm for general MAP estimation problems under linear Gaussian observation models. We also study the relationship between the CFLB/MM algorithm and the EM algorithm. We then modify the proposed algorithm to image denoising. We show that the proposed image denoising algorithm can be regarded as a generalization of the classical Wiener estimate algorithm. We propose a noniterative and an iterative algorithm for image denoising. We discuss connections of the proposed iterative algorithm with those algorithms using empirical Bayes and issues related to using the proposed algorithms in over-complete wavelet representations. Experimental results show that the performance of the proposed algorithm using a single wavelet representation is better than that of the bi-shrinkage algorithm which is arguably one of the best in recent publications. When over-complete wavelet representations such as the complex wavelets are used, the performance of the proposed algorithms are competitive with three state-of-the-art algorithms.

REFERENCES

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing Estimation Theory*, Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[2] P. J. Huber, *Robust Statistics*, John Wiley & Sons, New York, NY, USA, 1981.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2001.

[4] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, London, UK, 1995.

[5] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[6] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1150–1159, 2003.

[7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[8] R. Tibshirani, "Regression shrinkage and selection via lasso," *Journal of the Royal Statistical Society. Series B*, vol. 57, pp. 257–288, 1995.

[9] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 939–967, 2001.

[10] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[11] J. M. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: a GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 937–951, 2006.

[12] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B*, vol. 36, pp. 99–102, 1974.

[13] W. H. Rogers and J. W. Tukey, "Understanding some long-tailed distributions," *Statistica Neerlandica*, vol. 26, pp. 211–226, 1962.

[14] K. L. Lange and J. S. Sinsheimer, "Normal/independent distributions and their applications in robust regression," *Journal of Computational and Graphical Statistics*, vol. 2, no. 2, pp. 175–198, 1993.

[15] L. Chuanhai, "Bayesian robust multivariate linear regression with incomplete data," *Journal of the American Statistical Association*, vol. 91, no. 435, pp. 1219–1227, 1996.

[16] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, 2003.

[17] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[18] F. Girosi, "Models of noise and robust estimates," Tech. Rep. Memo 1287, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1991.

[19] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, Ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[20] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Advanced Mean Field Methods: Theory and Practice*, D. Saad and M. Opper, Eds., MIT Press, Cambridge, Mass, USA, 2000.

[21] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[22] M. J. Beal, *Variational algorithms for approximate Bayesian inference*, Ph.D. dissertation, The Gatsby Computational Neuroscience Unit, University College of London, London, UK, May 2003.

[23] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.

[24] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[25] M. Welling, G. E. Hinton, and S. Osindero, "Learning sparse topographic representations with products of student-t distributions," in *Proceedings of the 15th Conference on Neural Information Processing Systems (NIPS '02)*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds., pp. 1359–1366, MIT Press, Vancouver, BC, Canada, December 2002.

[26] A. C. Likas and N. P. Galatsanos, "A variational approach for Bayesian blind image deconvolution," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2222–2233, 2004.

[27] D. R. Hunter and R. Li, "Variable selection using MM algorithms," *Annals of Statistics*, vol. 33, no. 4, pp. 1617–1642, 2005.

[28] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *Journal of Computational & Graphical Statistics*, vol. 9, pp. 60–77, 2000.

[29] Z. Zhang, J. T. Kwok, and D.-Y. Yeung, "Surrogate maximization/minimization algorithms for AdaBoost and the logistic regression model," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 927–934, Banff, Canada, July 2004.

[30] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[31] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[32] M. A. T. Figueiredo and R. D. Nowak, "A bound optimization approach to wavelet-based image deconvolution," in *Proceedings of the International Conference on Image Processing (ICIP '05)*, vol. 2, pp. 782–785, Genova, Italy, September 2005.

[33] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira, "Total variation image deconvolution: a majorization-minimization approach," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 2, Toulouse, France, May 2006.

[34] G. Deng and W.-Y. Ng, "A minorization-maximization algorithm for maximum a posteriori signal estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 2, pp. 617–620, Toulouse, France, May 2006.

[35] K. Lange and J. Fessler, "Globally convergent algorithms for maximuma posteriori transmission tomography," *IEEE Transactions on Image Processing*, vol. 4, no. 10, pp. 1430–1438, 1995.

[36] A. de Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Transactions on Medical Imaging*, vol. 14, no. 1, pp. 132–137, 1995.

[37] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 298–311, 1997.

[38] H. Erdogan and J. A. Fessler, "Monotonic algorithms for transmission tomography," *IEEE Transactions on Medical Imaging*, vol. 18, no. 9, pp. 801–814, 1999.

[39] L. Şendur and I. W. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 438–441, 2002.

[40] X. Li and M. T. Orchard, "Spatially adaptive image denoising under overcomplete expansion," in *Proceedings of IEEE International Conference on Image Processing (ICIP '00)*, vol. 3, pp. 300–303, Vancouver, BC, Canada, September 2000.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[42] D. R. Hunter, K. Lange, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[43] M. Jamshidian, "Adaptive robust regression by using a nonlinear regression program," *Journal of Statistical Software*, vol. 4, pp. 1–25, 1999.

[44] G. Deng, "Generalized wiener estimation algorithms based on a family of heavy-tail distributions," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 1, pp. 457–460, Genova, Italy, September 2005.

[45] A. Gelman, H. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2004.

[46] G. Deng, "Signal estimation using multiple-wavelet representations and Gaussian models," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 1, pp. 453–456, Genova, Italy, September 2005.

[47] L. Şendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2744–2756, 2002.

[48] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

[49] M. K. Mihcak, I. Kozintsev, K. Ramchandram, and P. Moulin, "Low-complexity image denoising based on statistical modelling of wavelet coefficients," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 300–303, 1999.

[50] M. Elad, "Why simple shrinkage is still relevant for redundant representations?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5559–5569, 2006.

[51] S. S. Chen, *Basis pursuit*, Ph.D. dissertation, Department of Statistics, Stanford University, Stanford, Calif, USA, November 1995.