Why Simple Shrinkage is Still Relevant for Redundant Representations? *

Michael Elad
Department of Computer Science
The Technion - Israel Institute of Technology
Haifa 32000 Israel, email: elad@cs.technion.ac.il

June 30, 2006

Abstract

Shrinkage is a well known and appealing denoising technique, introduced originally by Donoho and Johnstone in 1994. The use of shrinkage for denoising is known to be optimal for Gaussian white noise, provided that the sparsity on the signal's representation is enforced using a unitary transform. Still, shrinkage is also practiced with non-unitary, and even redundant representations, typically leading to very satisfactory results. In this paper we shed some light on this behavior. The main argument in this paper is that such simple shrinkage could be interpreted as the first iteration of an algorithm that solves the basis pursuit denoising (BPDN) problem. While the desired solution of BPDN is hard to obtain in general, we develop in this paper a simple iterative procedure for the BPDN minimization that amounts to step-wise shrinkage. We demonstrate how the simple shrinkage emerges as the first iteration of this novel algorithm. Furthermore, we show how shrinkage can be iterated, turning into an effective algorithm that minimizes the BPDN via simple shrinkage steps, in order to further strengthen the denoising effect.

Keywords: Basis Pursuit, Overcomplete, Redundant, Frame, Sparse representation, Shrinkage, Denoising, Thresholding.

*This research was supported by the Technion's VPR funds, and the Gabriel and Matilda Brent Trust.

1 Introduction

A commonly practiced approach towards the removal of additive Gaussian white noise from a signal $\mathbf{y}$ is the minimization of the function

$$f(\mathbf{x}) = \frac{1}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \cdot P_r\{\mathbf{x}\} \ . \tag{1}$$

This expression can be shown to emerge from a Baysian point of view, when deploying the maximum a-posteriori probability (MAP) estimation [4]. The first term is commonly referred to as the log-likelihood, describing the relation between the desired (clean) signal, $\mathbf{x} \in \mathbb{R}^N$, and a noisy version of it $\mathbf{y} \in \mathbb{R}^N$. We assume the model $\mathbf{y} = \mathbf{x} + \mathbf{v}$, with $\mathbf{v} \in \mathbb{R}^N$ a Gaussian zero mean white noise[1]. The term $P_r\{\mathbf{x}\}$ stands for a prior posed on the unknown signal $\mathbf{x}$; numerous such expressions have been used for various signal types, as can be found in the literature. Among the popular methods that are now considered as classic methods in signal processing, we mention:

1. Energy minimization $P_r\{\mathbf{x}\} = \|\mathbf{x}\|_2^2$ – the simplest to handle, leading to $\hat{\mathbf{x}} = \mathbf{y}/(1+\lambda)$, which could be interpreted as a trivial shrinkage operation, as the signal is attenuated as part of the restoration;

2. Smoothness penalty $P_r\{\mathbf{x}\} = \|\mathbf{L}\mathbf{x}\|_2^2$ using the Laplacian operator, leading to the Wiener filter, $\hat{\mathbf{x}} = (\mathbf{I} + \lambda \mathbf{L}^T\mathbf{L})^{-1}\mathbf{y}$;

3. Total variation (TV) handles smoothness while allowing for sharp edges, $P_r\{\mathbf{x}\} = \||\nabla \mathbf{x}|\|_1$. This is done by an iterative update rule of the form $\hat{\mathbf{x}}_{k+1} = (1-\mu)\hat{\mathbf{x}}_k + \mu\mathbf{y} - \mu\lambda\nabla^T(\nabla\mathbf{x}_k/|\nabla\mathbf{x}_k|)$ [1];

4. Scalar entropy of $\mathbf{x}$, defined by $P_r\{\mathbf{x}\} = -\mathbf{x}^T\log(\mathbf{x})$. This promotes non-uniformity in $\mathbf{x}$, while assuming non-negative values only. Here again one cannot have a closed form solution, and an iterative procedure of the form $\hat{\mathbf{x}}_{k+1} = \exp\{\mathbf{1} - (\mathbf{x}_k - \mathbf{y})/\lambda\}$ can be used [2]; and

5. Sparsity of the unknown signal with respect to its transformed representation, $P_r\{\mathbf{x}\} = \|\mathbf{T}\mathbf{x}\|_1$. Again, iterative solver should be applied in general, giving an update of the form $\hat{\mathbf{x}}_{k+1} = (1-\mu)\hat{\mathbf{x}}_k + \mu\mathbf{y} - \mu\lambda \cdot \mathbf{T}^T\text{sign}(\mathbf{T}\mathbf{x}_k)$ [26].

[1] In most papers, when the noise characteristics are introduced, an accompanying "no-restriction" statement is added, suggesting that other models can be easily handled similarly. In this work this is not the case. The assumptions about the noise being Gaussian and white are crucial and cannot be replaced with alternatives such as colored noise or different distributions. In other words, shrinkage, as will be discussed here, is tightly coupled with these assumptions.

Formulations of the denoising problem as in (1) are commonly used, and are the basis for many more general inverse problems. Using priors such as TV and other robust methods, the above opens an opportunity to the study of non-linear filtering, typically formulated as partial differential equations (PDE). As we have seen in the above list (items #3 & #5), those are typically handled by iterative numerical solvers, that are characterized as slow and cumbersome[2].

In parallel to the progress made in the PDE directions, Donoho and Johnstone pioneered a wavelet based signal denoising algorithm in line with the above list of priors (item #5). In a sequence of publications they advocated the use of sparsity of the wavelet coefficients as a driving force in recovering the desired signal [5, 6, 7, 8, 9, 10]. Later work in [11, 12, 13] simplified these ideas and related them to the MAP formulation as presented above, using the prior $P_r\{\mathbf{x}\} = \|\mathbf{W}\mathbf{x}\|_p$, with a unitary wavelet transform matrix[3] $\mathbf{W} \in \mathbb{R}^{N\times N}$, and $0 \leq p \leq 1$.

Interestingly, using such a prior in Equation (1) leads to a *simple closed-form solution, known as shrinkage* (for the sake of completeness of the discussion, Section 2 establishes that). This solution amounts to a wavelet transform on the noisy signal, a look-up-table (LUT)[4] function on the coefficients (that depends on p), $\mathcal{S}\{\mathbf{W}\mathbf{y}\}$, and an inverse wavelet transform to produce the outcome $\hat{\mathbf{x}}$. Figure 1 presents this appealing and simple algorithm. Such a direct solution stands as a refreshing alternative to the iterative and slow methods mentioned above.

In Section 2 we follow [5, 6, 7, 8, 9, 10, 12, 13], and present the analysis that shows how the shrinkage algorithm becomes indeed the optimal solver of (1). This optimality depends strongly on the ℓ^2-norm used in evaluating the distance $\mathbf{x} - \mathbf{y}$, and this has direct roots in the white Gaussianity assumptions on the noise. Also, crucial to the optimality of this method is the orthogonality of $\mathbf{W}$.

A new trend of recent years is the use of overcomplete transforms, replacing the traditional unitary ones – see [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 29, 30, 31, 32, 33] for representative works. This

[2]Speedup algorithms have been proposed (e.g., the AOS method by Weickert [3] or the bilateral filter [4]), but those are still quite involved compared to simplicity of shrinkage – see it's description next.

[3]As will be shown hereafter, it is the unitarity of $\mathbf{W}$ that makes this choice so appealing, whereas the fact that this is specifically chosen as the wavelet transform has to do with the type of signals handled.

[4]LUT means a 1D memoryless function that operates on each wavelet coefficient in the same way. Assuming that the input is discretized to a finite number of bits, such operation can be done by considering the incoming value as an address to a pre-organized table of output values, and thus the name look-up-table.

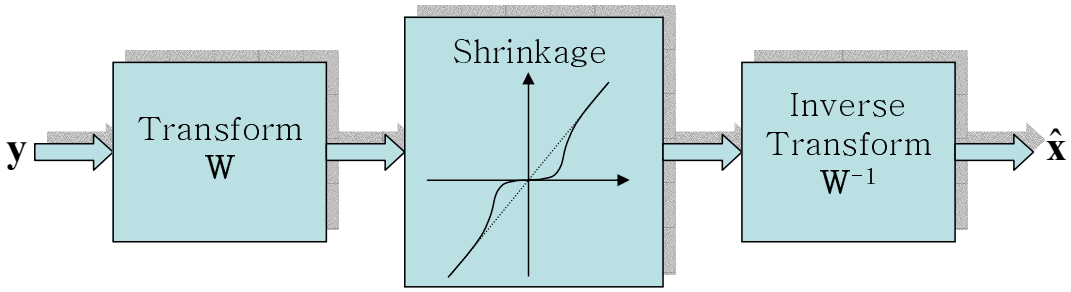

Figure 1: A block diagram of the shrinkage method for denoising.

trend was partly motivated by the growing realization that orthogonal wavelets are weak in describing the singularities found in images. The sources of this weakness are the loss of shift-invariance due to mandatory decimation, and the separability forced, which implies a lack of directional treatment. Another driving force in the introduction of redundant representations is the sparsity it can provide, which many applications find desirable. Finally, we should mention the desire to obtain shift-invariant transforms, again calling for redundancy in the representation.

In these methods the transform is defined via a non-square full rank matrix $\mathbf{T} \in \mathbb{R}^{L \times N}$, with $L > N$. Such redundant methods, like the un-decimated wavelet transform, curvelet, contourlet, steerable-wavelet, and more, were shown to be more effective in representing images, and other signal types. It is often assumed that $\mathbf{T}$ is a tight frame, implying that $\alpha \mathbf{T}^T \mathbf{T} = \mathbf{I}$. In such a case, the adjoint $\mathbf{T}^T$ stands for the Moore-Penrose pseudo-inverse transform, up to the constant α.

Given a noisy signal $\mathbf{y}$, one can still follow the shrinkage procedure, by computing the forward transform $\mathbf{Ty}$, putting the coefficients through a shrinkage LUT operation $\mathcal{S}\{\mathbf{Ty}\}$, and finally applying the inverse transform[5] to obtain the denoised outcome, $\mathbf{T}^{+}\mathcal{S}\{\mathbf{Ty}\}$. Will this be the solver of (1) when using the prior $P_r\{\mathbf{x}\} = \|\mathbf{Tx}\|_p$? The answer is negative. As we have said before, the orthogonality of the transform plays a crucial role in the construction of the shrinkage as an optimal procedure. Still, shrinkage is practiced quite often with non-unitary, and even redundant representations, typically leading to results better than in the non-redundant cases – see [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25] for representative examples. Naturally, we should wonder why this is so.

[5]There are infinitely many ways to define the inverse, and in most cases the Moore-Penrose pseudo-inverse is practiced.

In this paper we shed some light on this behavior. Our main argument, as will be composed in Section 3, is that such a shrinkage could be interpreted as the first iteration of a converging algorithm that solves the basis pursuit denoising (BPDN) problem [26]. The BPDN forms a similar problem to the one posed in (1), replacing the analysis prior with a generative one. While the desired solution of BPDN is hard to obtain in general, a simple iterative procedure that amounts to step-wise shrinkage can be employed with quite successful performance. Thus, beyond showing that shrinkage has justified roots in solid denoising methodology, we show how shrinkage can be iterated in a simple form, to further strengthen the denoising effect. As a byproduct, we get an effective algorithm that minimizes the BPDN functional via simple shrinkage steps.

This paper is organized as follows: In Section 2 we show how shrinkage emerges as an optimal solution for the prior $P_r\{\mathbf{x}\} = \|\mathbf{W}\mathbf{x}\|_p$ with *any* unitary matrix $\mathbf{W}$. This is a well-known result, belonging now to the classics of signal processing. We bring it here for completeness, and to set the stage for the redundant representation case, whose analysis follows. In Section 3 we generalize the prior to use redundant transforms. We show first the BPDN formulation as the desired denoising method, and then show how shrinkage could approximate it. In Section 4 we present some experimental results to support the claims made in Section 3.

2 Shrinkage For Unitary Transforms

In this Section we consider the denoising problem with a general additive prior that utilizes an orthonormal matrix $\mathbf{W}$. The ideas presented in this Section can be traced back to [5], and also found in [6, 7, 8, 9, 10, 11, 12, 13], although it may appear as posed somewhat differently. We intend to minimize

$$f(\mathbf{x}) = \frac{1}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{W}\mathbf{x}\} \quad . \tag{2}$$

In our notations, $\rho(\cdot)$ is an arbitrary scalar function. When applied to a vector, it is producing an output vector obtained by operating on each entry independently. The vector $\mathbf{1} \in \mathbb{R}^N$ is a vector of ones. Thus, the prior term amounts to

$$\mathbf{1}^T \cdot \rho\{\mathbf{W}\mathbf{x}\} = \sum_{n=1}^{N} \rho\{[\mathbf{W}\mathbf{x}]_n\}, \tag{3}$$

and thus the name "additive prior".

We typically assume (for convenience, and those assumptions can be relaxed) that $\rho(z)$ is symmetric ($\rho(z) = \rho(-z)$), and monotonic non-decreasing in the range $z > 0$, implying $\rho'(z) \geq 0$. As examples, choosing $\rho(z) = |z|^2$ leads to $P_r\{\mathbf{x}\} = \|\mathbf{W}\mathbf{x}\|_2^2$, choosing $\rho(z) = |z|$ gives the ℓ^1 alternative – $P_r\{\mathbf{x}\} = \|\mathbf{W}\mathbf{x}\|_1$, and $\rho(z) = |z|^p$ leads to the ℓ^p option – $P_r\{\mathbf{x}\} = \|\mathbf{W}\mathbf{x}\|_p^p$, all being special cases of this general additive form.

Defining $\mathbf{x}_w = \mathbf{W}\mathbf{x}$ and similarly $\mathbf{y}_w = \mathbf{W}\mathbf{y}$, the function $f(\mathbf{x})$ in (2) can be rearranged to become a function of $\mathbf{x}_w$,

$$f(\mathbf{x}_w) = \frac{1}{2} \cdot \|\mathbf{W}^{-1}(\mathbf{x}_w - \mathbf{y}_w)\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{x}_w\} \ . \tag{4}$$

Exploiting the unitary invariance property of the ℓ^2-norm[6], we can discard of the multiplication by $\mathbf{W}^{-1}$ in the first term, and obtain

$$\begin{aligned} f(\mathbf{x}_w) &= \frac{1}{2} \cdot \|(\mathbf{x}_w - \mathbf{y}_w)\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{x}_w\} = \\ &= \sum_{n=1}^{N} \left[\frac{1}{2} (x_w(n) - y_w(n))^2 + \lambda\rho(x_w(n)) \right] = f(x_w(1),\ x_w(2),\ \ldots\ , x_w(N)) \,. \end{aligned} \tag{5}$$

A consequence of the above simple manipulation is the fact that the original problem with respect to the unknown $\mathbf{x}_w$ is now decoupled. It can be solved independently for each unknown entry $x_w(n)$ as a scalar optimization procedure, and this is far easier than the N-dimensional optimization task we embarked from.

Note that the first stage to be done in solving (5) is to transform the input signal $\mathbf{y}$ to obtain $\mathbf{y}_w = \mathbf{W}\mathbf{y}$. This aligns with the block diagram described in Figure 1. Once this is done, we face N 1D optimization problems of the general form

$$z_{opt} = \arg\min_z g(z, a) = \arg\min_z \frac{1}{2} (z - a)^2 + \lambda\rho(z) \,, \tag{6}$$

with a and λ assumed known. The solution is an anti-symmetric LUT curve of the form $z_{opt} = \psi(a)$. Indeed, for an arbitrary a we have

$$g(z, -a) = \frac{1}{2} (z - (-a))^2 + \lambda\rho(z) = \frac{1}{2} ((-z) - a)^2 + \lambda\rho(-z) = g(-z, a),$$

where we have exploited the fact that ρ is symmetric. Thus, if for some $a > 0$ we have that $z_{opt} = \psi(a)$, then necessarily, $\psi(-a) = -z_{opt} = -\psi(a)$, being anti-symmetric as claimed. Thus, we can restrict our

[6]This is where we exploit both the orthogonality of $\mathbf{W}$ and the white Gaussianity of the noise, as promised.

analysis to positive inputs $a \geq 0$. Assuming that ρ is continuously differentiable[7], $\rho(\cdot) \in \mathbb{C}^1$, the solution should satisfy

$$z_{opt} = a - \lambda \rho' \left(z_{opt} \right). \tag{7}$$

This implicit equation can give the curve $\psi(a)$. An alternative approach to obtain $\psi(a)$ is via a direct numerical minimization of $g(z, a)$, and this can be done *even if* $\rho(\cdot)$ *is non-convex*, still leading to the global minimum solution of the penalty function given in Equation (2). Note that addressing this non-convex penalty term in Equation (2) using classical iterative optimizers (steepest descent, conjugate gradient, Newton methods, etc.) is susceptible to local minima traps in general, and depend on the initialization chosen. Here, not only iterations have been avoided, but finding of the best solution is guaranteed.

As can be seen, if we assume that $\rho(\cdot)$ is monotonic non-decreasing for $a \geq 0$, Equation (7) implies that the curve is sub-linear, giving a shrinkage of the input coefficient a, and thus the name given to this procedure. Figure 2 presents several such curves. Those where obtained by numerically finding the minimum of (6) for varying values of a.

After taking every coefficient $y_w(n)$ and passing it through the curve $\psi(y_w(n))$, the output are the representation coefficients of $\mathbf{x}$. A final stage of inverse transform provides the desired solution $\mathbf{x}$, as Figure 1 shows.

3 Treating Redundant Transforms

We turn now to the case where the prior term is given by $P_r \{\mathbf{x}\} = \mathbf{1}^T \cdot \rho(\mathbf{T}\mathbf{x})$, with $\mathbf{T}$ being a non-square full rank matrix $\mathbf{T} \in \mathbb{R}^{L \times N}$, where $L > N$. The discussion below will be mostly restricted to the choice $\rho(z) = |z|$, although all the discussion can be easily generalized to other choices of $\rho(\cdot)$.

Let us first describe the heuristic shrinkage algorithm that could be done for denoising a signal under these circumstances. This algorithm is described as *Algorithm A*.

[7] Or can be presented as the limit of a sequence of such functions, which is a more reasonable assumption.

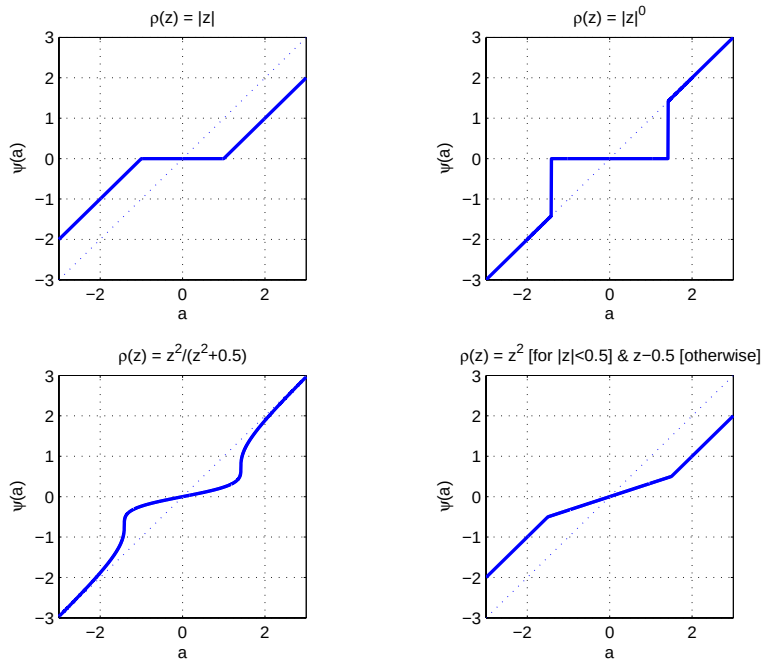

Figure 2: Several examples of the shrinkage LUT curve $\psi_\rho(a)$. In all these cases $\lambda = 1$.

Task: Denoise $\mathbf{y}$ using a heuristic shrinkage, based on the transform $\mathbf{T}$.

Data and Parameters: λ, $\mathbf{T}$, and $\mathbf{y}$.

Step 1: Compute $\mathbf{y}_T = \mathbf{T}\mathbf{y}$.

Step 2: Apply shrinkage $\mathcal{S}\{\mathbf{y}_T\}$ with threshold λ.

Step 3: Compute the inverse transform $\hat{\mathbf{x}} = \mathbf{T}^+\mathcal{S}\{\mathbf{y}_T\}$.

Finalize: The denoised output is $\hat{\mathbf{x}}$.

Algorithm A - Heuristic Shrinkage.

Note that in the above-described algorithm, it is unclear whether using a fixed and equal threshold to all coefficients is the proper way to go. We will see that this matter resolves as we relate this algorithm to a solid Baysian objective. We thus turn now to define the Bayesian objective for our denoising task.

3.1 The Generative Bayesian Objective

Following the preceding discussion, we intend to minimize a penalty similar to the one posed in (2), with an updated prior term:

$$f_1(\mathbf{x}) = \frac{1}{2} \cdot \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{T}\mathbf{x}\} \ . \tag{8}$$

Defining $\mathbf{x}_T = \mathbf{T}\mathbf{x}$, multiplying both sides by $\mathbf{T}^T$ we get $\mathbf{T}^T\mathbf{x}_T = \mathbf{T}^T\mathbf{T}\mathbf{x}$. In the general case, $\mathbf{T}^T\mathbf{T}$ is invertible (since $\mathbf{T}$ is full-rank), and thus[8] $\mathbf{x} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{x}_T = \mathbf{T}^+\mathbf{x}_T$. We can use these relations to rearrange Equation (8) and obtain a new function of the representation vector $\mathbf{x}_T$,

$$\begin{aligned} f_2(\mathbf{x}_T) &= \frac{1}{2} \cdot \|\mathbf{T}^+\mathbf{x}_T - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{x}_T\} \\ &= \frac{1}{2} \cdot \|\mathbf{D}\mathbf{x}_T - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{x}_T\}, \end{aligned} \tag{9}$$

where we denote $\mathbf{D} = \mathbf{T}^+$.

Denoising can be done by minimizing f_1 and obtaining a solution $\hat{\mathbf{x}}_1$. Alternatively, we can minimize f_2 with respect to $\mathbf{x}_T$ and deduce the denoised outcome by $\hat{\mathbf{x}}_2 = \mathbf{D}\hat{\mathbf{x}}_T$. Interestingly, *these two results are not expected to be the same in the general case*, since in the conversion from f_1 to f_2 we have expanded the set of feasible solutions by allowing $\mathbf{x}_T$ to be an arbitrary vector in $\mathbb{R}^L$, whereas the original definition $\mathbf{x}_T = \mathbf{T}\mathbf{x}$ implies that it must be confined to the column space of $\mathbf{T}$. Notice that this difference between the two formulations disappears when $\mathbf{T}$ is full rank square matrix, which explains why this dichotomy of methods did not bother us in the previous section.

Still, the formulation posed in (9) is a feasible alternative Bayesian method that uses a generative prior. Indeed, for the choice $\rho\{z\} = |z|$, this formulation is known as the basis pursuit denoising (BPDN). Referring to $\mathbf{D}$ as a dictionary of signal prototypes (atoms) as its columns, we assume that the desired signal $\mathbf{x}$ is a linear construction of these atoms, with coefficients drawn independently from a probability density function proportional to $\exp\{-\text{Const} \cdot \rho\{x_T(j)\}\}$. In the case of $\rho(z) = |z|$ this is the Laplace distribution, and we effectively promote sparsity in the representation.

[8] If $\mathbf{T}$ is a tight frame ($\alpha\mathbf{T}^T\mathbf{T} = \mathbf{I}$), then the above leads to $\mathbf{x} = \alpha\mathbf{T}^T\mathbf{x}_T$.

3.2 Our Objective

Our objective is now the optimization problem

$$\min_{\mathbf{z}} \frac{1}{2} \cdot \|\mathbf{D}\mathbf{z} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{z}\}, \tag{10}$$

and the denoised output is $\hat{\mathbf{x}} = \mathbf{D}\hat{\mathbf{z}}$. Although BPDN has been defined for the specific choice $\rho(z) = |z|$, we shall refer hereafter to this more general objective as BPDN as well.

The numerical method we shall use in later experiments to compare against shrinkage is based on the Iterative Reweighed Least-Squares (IRLS) as practiced by the FOCUSS algorithm [27, 28]. We outline it here very briefly: Define a new scalar function $\rho_0(z)$ to satisfy $\rho(z) = 0.5|z|^2 \cdot \rho_0(z)$. For example, for $\rho(z) = |z|$, we have $\rho_0(z) = 2/|z|$ (or better yet, $\rho_0(z) = 2/(|z| + \epsilon)$ with $0 < \epsilon \ll 1$ so as to avoid divisions by zero). Now, the above formulation can be rewritten as

$$\min_{\mathbf{z}} \frac{1}{2} \cdot \|\mathbf{D}\mathbf{z} - \mathbf{y}\|_2^2 + \frac{\lambda}{2} \cdot \mathbf{z}^T \cdot \mathrm{diag}\{\rho_0\{\mathbf{z}\}\}\mathbf{z}. \tag{11}$$

The term $\mathrm{diag}\{\rho_0\{\mathbf{z}\}\}$ is a diagonal weight matrix, and thus the second term has a quadratic structure, if we assume this diagonal matrix to be fixed. Minimization can be done iteratively by assuming that the term $\rho_0\{\mathbf{z}\}$ is fixed, being computed with the current solution $\mathbf{z}_k$. Then the problem has a simple quadratic form, leading to the update equation

$$\mathbf{z}_{k+1} = \left[\mathbf{D}^T\mathbf{D} + \lambda \mathrm{diag}\{\rho_0\{\mathbf{z}_k\}\}\right]^{-1}\mathbf{D}^T\mathbf{y}. \tag{12}$$

The IRLS algorithm is described as *Algorithm B*.

Task: Denoise the signal $\mathbf{y}$ by $\hat{\mathbf{x}} = \arg\min_{\mathbf{z}} \frac{1}{2} \cdot \|\mathbf{D}\mathbf{z} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \|\mathbf{z}\|_1$.

Data and Parameters: λ, $\mathbf{D}$, and $\mathbf{y}$ are given, K_0 iterations.

Initialization: Set $k = 0$ and choose $\mathbf{z}_k = \mathbf{1}$.

Main Iteration: Set $k = 1$ and apply:

- **Weights:** Compute $\mathbf{W} = \mathrm{diag}\{\rho_0\{\mathbf{z}_{k-1}\}\} = 2\mathrm{diag}\{1/|\mathbf{z}_{k-1}|\}$.
- **Update:** Compute $\mathbf{z}_k = \left[\mathbf{D}^T\mathbf{D} + \lambda\mathbf{W}\right]^{-1}\mathbf{D}^T\mathbf{y}$.
- **Return:** Set $k = k + 1$. If $k \leq K_0$ return to "Weights".

Finalize: The denoised output is $\hat{\mathbf{x}} = \mathbf{D}\mathbf{z}_{K_0}$.

Algorithm B - The Iterative Reweighed Least-Squares.

Notice that we have recommended an initialization with ones. Using a zero initialization causes a slow start because then $\mathbf{D}^T\mathbf{D}$ is negligible compared to $\mathbf{W}$. The choice of ones parallels a regularized pseudo-inverse start.

In the experiments to follow we shall implement this algorithm to minimize the function in (10), but it should be clear that in general this is a daunting task for typical sizes used in image processing ($N \approx 10e+4$, $L \approx 10e+6$, and beyond). The need to invert a matrix of size $L \times L$, as the above update formula requires, is prohibitive, and should be replaced with an iterative solver. In fact, this is why a shrinkage method as in *Algorithm A* would be of interest in the first place.

3.3 Shrinkage Again? A Sequential Method

For a dictionary built as a union of J ortho-matrices, there is yet another, more efficient, numerical solver, based on a block-coordinate-relaxation (BCR) process. This algorithm, as introduced by Bruce et. al. [34] is using shrinkage in the spirit of Section 2. The representation vector $\mathbf{z}$ is broken into J parts, each referring to a unitary matrix in $\mathbf{D}$. The BCR algorithm addresses one set of representation coefficients at a time, assuming all the others as fixed. We will imitate this idea here, but consider a general dictionary, and treat scalar entries in $\mathbf{z}$.

Assume that in an iterative process used to solve the above problem, we hold the k-th solution $\hat{\mathbf{z}}_k$. We are interested in updating its j-th entry, $z(j)$, assuming all the others as fixed. Thus, we obtain a one-dimensional optimization problem of the form

$$\min_{w} \frac{1}{2} \cdot \| \left[\mathbf{D}\mathbf{z}_k - \mathbf{d}_j z_k(j)\right] + \mathbf{d}_j w - \mathbf{y}\|_2^2 + \lambda \cdot \rho\left\{w\right\}. \tag{13}$$

In the above expression, $\mathbf{d}_j$ is the j-th column in $\mathbf{D}$. The term $\mathbf{D}\mathbf{z}_k - \mathbf{d}_j z_k(j)$ uses the current solution for all the coefficients, but discards of the j-th one, assumed to be replaced with a new value, w.

Since this is a 1D optimization task, it is relatively easy to solve. Assume for example that $\rho(w) = |w|$ as was done in the previous Section. Taking a derivative with respect to w we get the equation

$$0 = \mathbf{d}_j^T \left(\left[\mathbf{D}\mathbf{z}_k - \mathbf{d}_j z_k(j)\right] + \mathbf{d}_j w - \mathbf{y}\right) + \lambda \cdot \text{sign}\left\{w\right\}, \tag{14}$$

leading to

$$w = \frac{\mathbf{d}_j^T \left(\mathbf{y} - \left[\mathbf{D}\mathbf{z}_k - \mathbf{d}_j z_k(j)\right]\right)}{\|\mathbf{d}_j\|_2^2} - \frac{\lambda \cdot \text{sign}\left\{w\right\}}{\|\mathbf{d}_j\|_2^2} = \tag{15}$$

$$
\begin{aligned}
&= z_k(j) + \frac{\mathbf{d}_j^T(\mathbf{y} - \mathbf{D}\mathbf{z}_k)}{\|\mathbf{d}_j\|_2^2} - \frac{\lambda \cdot \text{sign}\{w\}}{\|\mathbf{d}_j\|_2^2} \\
&= v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) - \hat{\lambda}(j) \cdot \text{sign}\{w\}.
\end{aligned}
$$

Here we have defined

$$
v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) = \frac{\mathbf{d}_j^T(\mathbf{y} - \mathbf{D}\mathbf{z}_k)}{\|\mathbf{d}_j\|_2^2} + z_k(j) \quad \text{and} \quad \hat{\lambda}(j) = \frac{\lambda}{\|\mathbf{d}_j\|_2^2}. \tag{16}
$$

Both $v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j)$ and $\hat{\lambda}(j)$ are computable using the known dictionary, noisy signal, current solution, the value of λ, and the index in question. Thus, the same reasoning as the one practiced in Section 2 leads to a closed form formula for the optimal solution for w, being a shrinkage operation on $v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j)$,

$$
w_{opt} = \mathcal{S}\{v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j)\} = \begin{cases} v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) - \hat{\lambda}(j) & \text{for } v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) > \hat{\lambda}(j) \\ 0 & \text{for } |v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j)| \leq \hat{\lambda}(j) \\ v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) + \hat{\lambda}(j) & \text{for } v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) < -\hat{\lambda}(j) \end{cases}. \tag{17}
$$

A similar LUT result can be developed for any other choice of the function $\rho(\cdot)$.

It is tempting to suggest an algorithm that applies the above procedure for $j = 1, 2, \ \ldots, \ , L$, updating one coefficient at a time in a sequential coordinate descent algorithm, and cycle such process several times. Such algorithm is described as *Algorithm C*.

Task: Denoise the signal $\mathbf{y}$ by $\hat{\mathbf{x}} = \arg\min_{\mathbf{z}} \frac{1}{2} \cdot \|\mathbf{D}\mathbf{z} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \|\mathbf{z}\|_1$.

Data and Parameters: λ, $\mathbf{D}$, and $\mathbf{y}$ are given, K_0 iterations.

Initialization: Set $k = 0$ and choose $\mathbf{z}_k = \mathbf{0}$.

Main Iteration: Set $k = 1$ and apply:

- **Sweep:** Set $j = 1$.
  1. Compute $v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j) = \frac{\mathbf{d}_j^T(\mathbf{y} - \mathbf{D}\mathbf{z}_k)}{\|\mathbf{d}_j\|_2^2} + z_k(j)$.
  2. Compute $w_{opt} = \mathcal{S}\{v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j)\}$ using threshold $\lambda / \|\mathbf{d}_j\|_2^2$.
  3. Update the solution at the j-th location to be $z_k(j) = w_{opt}$.
  4. Set $j = j + 1$. If $j \leq L$, return to step 1.
- **Return:** Set $k = k + 1$. If $k \leq K_0$ return to "Sweep".

Finalize: The denoised output is $\hat{\mathbf{x}} = \mathbf{D}\mathbf{z}_{K_0}$.

Algorithm C - Sequential Shrinkage.

While such algorithm necessarily converges, and could be effective in minimizing the objective function using scalar shrinkage operations only, it is impractical in most cases. The reason is the necessity to

draw one column at a time from $\mathbf{D}$ to perform this computation. Consider, for example, the curvelet dictionary. While the transform and its inverse can be interpreted as multiplications by the dictionary and its transpose (because it is a tight frame), this matrix is never explicitly constructed, and an attempt to draw basis functions from it or store them could be devastating. Thus we take a different route.

3.4 Shrinkage in a Parallel Method

Given the current solution $\mathbf{z}_k$, let us assume that we use the above update formulation to update *all the coefficients* in parallel, rather than doing this sequentially. Obviously, this process must be slower in minimizing the objective function, but with this slowness comes a much desired simplicity that will be evident shortly.

First, let us convert the terms $v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, j)$ in Equation (16) to a vector form that accounts for all the updates at once. Gathering these terms for all $j \in [1, L]$, this reads

$$\mathbf{v}(\mathbf{D}, \mathbf{y}, \mathbf{z}_k) = \begin{bmatrix} v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, 1) \\ v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, 2) \\ \vdots \\ v(\mathbf{D}, \mathbf{y}, \mathbf{z}_k, L) \end{bmatrix} = \text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}\mathbf{D}^T\left(\mathbf{y} - \mathbf{D}\mathbf{z}_k\right) + \mathbf{z}_k. \tag{18}$$

Notice that in the computation of $\mathbf{v}$ in the above equation, we do not need to extract some columns from the dictionary, and need not use these matrices explicitly in any other way. If the transform we use is such that multiplication by $\mathbf{D}$ and its adjoint $\mathbf{D}^T$ are fast, then computing the above term is easy and efficient. The normalization by the norms of the columns is simple to obtain and can be kept as fixed parameters of the transform, computed once off-line[9].

In the case of tight frames, applying multiplications by $\mathbf{D}^T$ and $\mathbf{D}$ are the forward and the inverse transforms, up to a constant. For a non-tight frame, the above formula says that we need to be able to apply the adjoint *and not the pseudo-inverse* of $\mathbf{D}$.

There is also a natural weakness to the above strategy. One cannot take a shrinkage of the above vector with respect to the threshold vector $\lambda \cdot \text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\} \cdot \mathbf{1}$, and expect the objective function to be minimized well. While updating every scalar entry w_j using the above shrinkage formula is necessarily decreasing the function's value, applying all those at once is likely to diverge, and cause an ascent in

[9]While storing the columns of $\mathbf{D}$ may require huge memory volume, storing a scalar per column is reasonable.

the objective. Thus, instead of applying a complete shrinkage as Equation (17) suggests, we consider a relaxed step of the form

$$\mathbf{z}_{k+1} = \mathbf{z}_k + \mu \left[\mathcal{S} \left\{ \mathbf{v}(\mathbf{D}, \mathbf{y}, \mathbf{z}_k) \right\} - \mathbf{z}_k \right] = \mathbf{z}_k + \mu \mathbf{h}_k. \tag{19}$$

This way, we compute the shrinkage vector as the formula suggests, and use it to define a descent direction. The solution is starting from the current solution $\mathbf{z}_k$ and updates it by "walking" towards the shrinkage result. For $\mu = 1$ the shrinkage is adopted in full, and for $\mu < 1$ the effect is a relaxed step. For a sufficiently small $\mu > 0$, this step *must* lead to a feasible descent in the penalty function, because this direction is a non-negative combination of L descent directions.

We can apply a line search to find the proper choice for the value of μ. In general, line search seeks the best step-size as a 1D optimization procedure that solves

$$\min_{\mu} \frac{1}{2} \cdot \| \mathbf{D} \left[\mathbf{z}_k + \mu \mathbf{h}_k \right] - \mathbf{y} \|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho \left\{ \mathbf{z}_k + \mu \mathbf{h}_k \right\}, \tag{20}$$

where $\mathbf{h}_k$ is a computable vector. The solution is given by solving the equation

$$0 = \mathbf{h}_k^T \mathbf{D}^T \left(\mathbf{D} \left[\mathbf{z}_k + \mu \mathbf{h}_k \right] - \mathbf{y} \right) + \lambda \cdot \mathbf{h}_k^T \cdot \rho' \left\{ \mathbf{z}_k + \mu \mathbf{h}_k \right\}. \tag{21}$$

This again has a shrinkage-like structure. Finding an appropriate μ amounts to a multiplication by $\mathbf{D}$ and its adjoint to compute the first term, and then seek optimal solution for μ by a zero-crossing iterative solver.

To summarize our findings so far, we desire a solution to the optimization problem posed in Equation (10). We do this iteratively, and update the result by performing a shrinkage. Defining the solution at the k-th iteration by $\mathbf{z}_k$, it is updated by computing $\mathbf{z}_k + \text{diag}^{-1} \left\{ \mathbf{D}^T \mathbf{D} \right\} \mathbf{D}^T \left(\mathbf{y} - \mathbf{D}\mathbf{z}_k \right)$, applying shrinkage to it using the threshold vector $\lambda \cdot \text{diag}^{-1} \left\{ \mathbf{D}^T \mathbf{D} \right\} \cdot \mathbf{1}$, and finally applying a line search on the line that connect $\mathbf{z}_k$ and the new result to get the best descent. This algorithm is described as *Algorithm D*.

Task: Denoise the signal $\mathbf{y}$ by $\hat{\mathbf{x}} = \arg\min_{\mathbf{z}} \frac{1}{2} \cdot \|\mathbf{Dz} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \|\mathbf{z}\|_1$.

Data and Parameters: λ, $\mathbf{D}$, $\mathbf{y}$, and K_0 (iterations) are given. $\mathbf{W} = \text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}$ is computed off-line.

Initialization: Set $k = 0$ and choose $\mathbf{z}_k = \mathbf{0}$.

Main Iteration: Set $k = 1$ and apply:

- **Error:** Compute $\mathbf{e} = \mathbf{y} - \mathbf{D}\mathbf{z}_{k-1}$.
- **Project:** Compute $\mathbf{e}_T = \mathbf{W}\mathbf{D}^T\mathbf{e}$.
- **Shrinkage:** Compute $\mathbf{e}_T^S = \mathcal{S}\left\{\mathbf{e}_T + \mathbf{z}_{k-1}\right\}$ with threshold $\lambda \cdot \mathbf{W} \cdot \mathbf{1}$.
- **Line-Search:** Find μ_0 to obtains a descent with $\mathbf{z}_{k-1} + \mu(\mathbf{e}_T^S - \mathbf{z}_{k-1})$.
- **Relax:** Update $\mathbf{z}_k = \mathbf{z}_{k-1} + \mu_0(\mathbf{e}_T^S - \mathbf{z}_{k-1})$.
- **Return:** Set $k = k + 1$. If $k \leq K_0$ return to "Error".

Finalize: The denoised output is $\hat{\mathbf{x}} = \mathbf{D}\mathbf{z}_{K_0}$.

Algorithm D - Parallel Shrinkage.

3.5 The First Iteration - A Closer Look

Let us now concentrate on the first iteration, assuming that the algorithm is initialized with $\mathbf{z}_0 = \mathbf{0}$. The term in Equation (18) becomes

$$\mathbf{v}(\mathbf{D}, \mathbf{y}, \mathbf{0}) = \text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}\mathbf{D}^T\mathbf{y}. \tag{22}$$

The solution $\mathbf{z}_1$ is obtained by first applying shrinkage to the above vector, using $\lambda\text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}\mathbf{1}$ as the threshold vector, and then relaxing it, as in Equation (19), giving

$$\mathbf{z}_1 = \mu\mathcal{S}\left\{\text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}\mathbf{D}^T\mathbf{y}\right\}. \tag{23}$$

As to the line-search for choosing μ, it is found by solving Equation (21). For the case of $\rho(z) = |z|$, this gives

$$\begin{aligned} 0 &= \mathbf{h}_1^T\mathbf{D}^T\left(\mu\mathbf{D}\mathbf{h}_1 - \mathbf{y}\right) + \lambda \cdot \mathbf{h}_1^T \cdot \rho'\left\{\mu\mathbf{h}_1\right\} \qquad (24)\\ &= \mathbf{h}_1^T\mathbf{D}^T\left(\mu\mathbf{D}\mathbf{h}_1 - \mathbf{y}\right) + \lambda\mu\mathbf{h}_1^T \cdot \text{sign}\left\{\mathbf{h}_1\right\} \\ &= \mu\|\mathbf{D}\mathbf{h}_1\|_2^2 - \mathbf{y}^T\mathbf{D}\mathbf{h}_1 + \lambda\mu\|\mathbf{h}_1\|_1. \end{aligned}$$

Thus, we choose

$$\mu = \frac{\mathbf{y}^T\mathbf{D}\mathbf{h}_1}{\|\mathbf{D}\mathbf{h}_1\|_2^2 + \lambda\|\mathbf{h}_1\|_1}, \tag{25}$$

and this can be computed by applying the multiplication by $\mathbf{D}$ only once more. To conclude, the denoised result is obtained by computing

$$\hat{\mathbf{x}} = \mu \cdot \mathbf{D}\mathcal{S}\left\{\text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}\mathbf{D}^T\mathbf{y}\right\}, \tag{26}$$

where the threshold to use in the shrinkage is given by $\lambda \cdot \text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\}\mathbf{1}$.

As a side note we mention that we can give a rough estimate to the value of μ under several simplifying assumptions. We assume that (i) $\mathbf{D}$ represents a tight frame, i.e., $\mathbf{D}\mathbf{D}^T = \alpha\mathbf{I}$; (ii) its columns are ℓ^2-normalized, i.e., $\text{diag}^{-1}\left\{\mathbf{D}^T\mathbf{D}\right\} = \mathbf{I}$; and (iii) the shrinkage operation is nearly transparent, i.e. $\mathcal{S}\left\{\mathbf{D}^T\mathbf{y}\right\} \approx \mathbf{D}^T\mathbf{y}$, meaning that the shrinkage almost does not change the data fed to it. Such is the case, for example, for a dictionary built as a union of ortho-matrices. Under these assumptions we obtain

$$\mu \approx \frac{\mathbf{y}^T\mathbf{D}\mathbf{D}^T\mathbf{y}}{\|\mathbf{D}\mathbf{D}^T\mathbf{y}\|_2^2 + \lambda\|\mathbf{D}^T\mathbf{y}\|_1} = \frac{\alpha \cdot \|\mathbf{y}\|_2^2}{\alpha^2 \cdot \|\mathbf{y}\|_2^2 + \lambda\|\mathbf{D}^T\mathbf{y}\|_1} \approx \frac{1}{\alpha}. \tag{27}$$

3.6 Relation to Heuristic Shrinkage

How all this compares with the heuristic shrinkage we described in *Algorithm A*? Recall that we have defined $\mathbf{D} = \mathbf{T}^+ = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$, and thus $\mathbf{D}^T = \mathbf{T} \cdot (\mathbf{T}^T\mathbf{T})^{-1}$. Going back to the final formula given in Equation (26) and using the relation between $\mathbf{T}$ and $\mathbf{D}$, we have that the denoising obtained here amounts to

$$\begin{aligned}\hat{\mathbf{x}} &= \mu \cdot \mathbf{T}^+\mathcal{S}\left\{\text{diag}^{-1}\left\{\left[\mathbf{T}^+\right]^T\mathbf{T}^+\right\}\left[\mathbf{T}^+\right]^T\mathbf{y}\right\} \\ &= \mu \cdot (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathcal{S}\left\{\text{diag}^{-1}\left\{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-2}\mathbf{T}^T\right\}\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{y}\right\}.\end{aligned} \tag{28}$$

For a general redundant transform, this formulation seems different from the one proposed in *Algorithm A* is several ways:

- Instead of starting by applying the transform on the signal, $\mathbf{T}\mathbf{y}$, the transposed inverse is applied, $\left[\mathbf{T}^+\right]^T\mathbf{y}$.
- The outcome is scaled element-by-element, by the matrix $\text{diag}^{-1}\left\{\left[\mathbf{T}^+\right]^T\mathbf{T}^+\right\}$, which does not appear in the heuristic algorithm.
- The shrinkage is applied with respect to a varying threshold, depending on the same diagonal matrix mentioned above. Thus, the threshold comparison is independent of this scale.

- In the return to the signal domain, both methods employ a multiplication by $\mathbf{T}^+$, however, here we have also a scale μ to take into account.

Are these algorithms truly so different? Let us consider several special cases.

Case 1: $\mathbf{T}$ is a tight frame, built with normalized columns of the dictionary. In this case $\alpha\mathbf{T}^T\mathbf{T} = \mathbf{I}$ and $\text{diag}^{-1}\left\{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-2}\mathbf{T}^T\right\} = \mathbf{I}$. Thus, Equation (28) becomes

$$\begin{aligned}\hat{\mathbf{x}} &= \mu\cdot(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathcal{S}\left\{\text{diag}^{-1}\left\{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-2}\mathbf{T}^T\right\}\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{y}\right\} \\ &= \mu\mathbf{T}^+\mathcal{S}\left\{\alpha\mathbf{T}\mathbf{y}\right\}.\end{aligned} \tag{29}$$

The shrinkage in this case is done with a threshold λ. Assuming $\rho(z) = |z|$, we have

$$\mathcal{S}\left\{\alpha[\mathbf{T}\mathbf{y}]_j\right\} = \begin{cases}\alpha[\mathbf{T}\mathbf{y}]_j - \lambda & \text{for } [\mathbf{T}\mathbf{y}]_j > \lambda/\alpha \\ 0 & \text{for } |[\mathbf{T}\mathbf{y}]_j| \le \lambda/\alpha \\ \alpha[\mathbf{T}\mathbf{y}]_j + \lambda & \text{for } [\mathbf{T}\mathbf{y}]_j < -\lambda/\alpha\end{cases}. \tag{30}$$

If we choose $\mu = 1/\alpha$ as previously suggested, this scales the above equation to give

$$\mu\mathcal{S}\left\{\alpha[\mathbf{T}\mathbf{y}]_j\right\} = \frac{1}{\alpha}\mathcal{S}\left\{\alpha[\mathbf{T}\mathbf{y}]_j\right\} = \begin{cases}[\mathbf{T}\mathbf{y}]_j - \lambda/\alpha & \text{for } [\mathbf{T}\mathbf{y}]_j > \lambda/\alpha \\ 0 & \text{for } |[\mathbf{T}\mathbf{y}]_j| \le \lambda/\alpha \\ [\mathbf{T}\mathbf{y}]_j + \lambda/\alpha & \text{for } [\mathbf{T}\mathbf{y}]_j < -\lambda/\alpha\end{cases}. \tag{31}$$

and this is the very same algorithm we presented in *Algorithm A* with a threshold chosen as λ/α.

Case 2: $\mathbf{T}$ is a general frame with normalized columns of the dictionary. In this case $\text{diag}^{-1}\left\{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-2}\mathbf{T}^T\right\} = \mathbf{I}$. Thus, Equation (28) becomes

$$\hat{\mathbf{x}} = \mu\mathbf{T}^+\mathcal{S}\left\{[\mathbf{T}^+]^T\mathbf{y}\right\}. \tag{32}$$

Given the matrix $\mathbf{T}$, we can define a new forward transform as $\tilde{\mathbf{T}}\mathbf{y} = [\mathbf{T}^+]^T\mathbf{y}$. With this definition, the above operation amounts to an application of the newly defined transform, application of shrinkage with a constant threshold λ, and then using the transform's adjoint to return to the signal domain. This stands as an interesting alternative to *Algorithm A*, being similar in some respects, and different in others. Specifically, we need to redefine how the transform operates, and we do not use its inverse but adjoint.

Case 3: $\mathbf{T}$ is a tight frame, with non-normalized columns. In this case $\alpha\mathbf{T}^T\mathbf{T} = \mathbf{I}$, and thus Equation (28) becomes

$$\hat{\mathbf{x}} = \mu\cdot(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathcal{S}\left\{\text{diag}^{-1}\left\{\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-2}\mathbf{T}^T\right\}\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{y}\right\} \tag{33}$$

$$= \quad \mu\alpha \cdot \mathbf{T}^T \mathcal{S}\left\{\frac{1}{\alpha}\text{diag}^{-1}\left\{\mathbf{T}\mathbf{T}^T\right\}\mathbf{T}\mathbf{y}\right\}.$$

The shrinkage in this case is done for the j-th entry with a threshold $\frac{\lambda}{\alpha^2}\|\mathbf{t}_j\|_2^{-2}$, where $\mathbf{t}_i$ is the j-th row in $\mathbf{T}$. Assuming $\rho(z) = |z|$, we have

$$\mathcal{S}\left\{\frac{[\mathbf{T}\mathbf{y}]_j}{\alpha\|\mathbf{t}_j\|_2^2}\right\} = \begin{cases} \frac{[\mathbf{T}\mathbf{y}]_j - \lambda/\alpha}{\alpha\|\mathbf{t}_j\|_2^2} & \text{for } [\mathbf{T}\mathbf{y}]_j > \lambda/\alpha \\ 0 & \text{for } |[\mathbf{T}\mathbf{y}]_j| \leq \lambda/\alpha \\ \frac{[\mathbf{T}\mathbf{y}]_j + \lambda/\alpha}{\alpha\|\mathbf{t}_j\|_2^2} & \text{for } [\mathbf{T}\mathbf{y}]_j < -\lambda/\alpha \end{cases}. \tag{34}$$

The multiplication after the shrinkage by $\alpha\mathbf{T}^T$ stands for a regular inverse transform, and thus the multiplication by μ should be absorbed in the shrinkage. This way we obtained a very similar algorithm to the one presented in *Algorithm A*, with a regular forward transform, regular shrinkage with a constant threshold λ/α, a scale of each entry by $\frac{\mu}{\alpha\|\mathbf{t}_j\|_2^2}$, and finally an inverse transform. Thus, the only difference is the element-wise scale of the transformed coefficients, prior to the return to the signal domain.

3.7 Discussion

The fact that our algorithm differs from *Algorithm A* for non-tight frames or non-normalized dictionaries should not be interpreted as a statement that *Algorithm A* is wrong. It may well be that an alternative justification could be developed, leading to *Algorithm A* as a first iteration. As an example, one could consider $\mathbf{D}^+\mathbf{y}$ as an initialization to our denoising process, based on the replacement of the ℓ^1-norm with an ℓ^2 one. Using this initialization, followed by appropriate adjustments after one iteration of some algorithm, it may be possible to give rise to a different shrinkage algorithm than the one we have developed. In this work we have not followed this line of reasoning.

An immediate benefit that can be drawn from the above results is the ability to operate the heuristic shrinkage in a better setting. Also, if we are willing to invest more computations, the above results give us a way to further minimize the objective by more iterations that are shrinkage-like, and this way possibly get stronger noise removal.

Also, from a different perspective, since BPDN is considered as an important objective functional (e.g., as a non-linear transform that promotes sparsity), what we have obtained here (as described in *Algorithm D*) could be an effective solver that uses only simple and fast operations. Thus, when applying a complicated transform such as curvelet or contourlet, instead of using the ℓ^2-based linear method for

obtaining the forward redundant transform, one can use the basis pursuit (or BPDN) and get a sparser outcome (see [45] for such experiments on image denoising using the BPDN with contourlet). *Algorithm D* can perform this task by applying the regular forward and adjoint transforms, coupled with simple shrinkage steps. As such, this algorithm can be perceived as a novel and effective pursuit algorithm.

4 Experimental Results

In this Section we present four experiments – the first three match the cases discussed in Section 3.6, illustrating the performance of the various algorithms discussed on some signal examples. The fourth experiment present the average denoising behavior of *Algorithm D* as a function of the noise power and the cardinality of the representation.

Experiment 1 – A tight frame with normalized columns: We build $\mathbf{D}$ as a union of 10 random unitary matrices of size 100×100. We synthesize a sparse representation $\mathbf{z}_0$ with 15 non-zeros in random locations and Gaussian i.i.d. entries, so as to match the sparsity prior we use. Thus the clean signal is defined as $\mathbf{x}_0 = \mathbf{D}\mathbf{z}_0$. This signal is contaminated by a Gaussian i.i.d noise $\sigma = 0.3$ (parallels an SNR of ≈ 1.3dB). We apply algorithms A-D with $\rho(z) = |z|$, and the results are reported in Figures 3-5.

First, we show how effective are algorithms B-D in minimizing the objective in Equation (10). Figure 3 presents the value of the objective as a function of the iteration number. Here we have implemented algorithm D both with a fixed $\mu = 1/\alpha$ and with a line-search. The IRLS (*Algorithm B*) performs the best in terms of convergence speed at the first 5 iterations, but then slows dramatically. The sequential and the parallel (with line-search) coordinate descent are comparable to each other, being somewhat inferior to the IRLS at the first 5 iterations, but show a consistent decent in the function value afterwards.

When implementing algorithms A-D, we sweep through the possible values of λ to find the best choice. In this paper we have not treated the question of how to automatically find it. Also, in assessing the denoising effect, we use the noise decay factor measure, $r(\hat{\mathbf{x}}, \mathbf{x}_0, \mathbf{y}) = \|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2 / \|\mathbf{y} - \mathbf{x}_0\|_2^2$, which gives the ratio between the final reconstruction error and the error with $\mathbf{y}$ as our estimate. Thus, a value smaller than 1 implies a decay in the noise, and the closer it is to zero the better the result.

Using the IRLS (with few iterations due to its fast convergence) can give us an evaluation of the denoising potential that exists in the objective function we use. We compare the IRLS results (after the

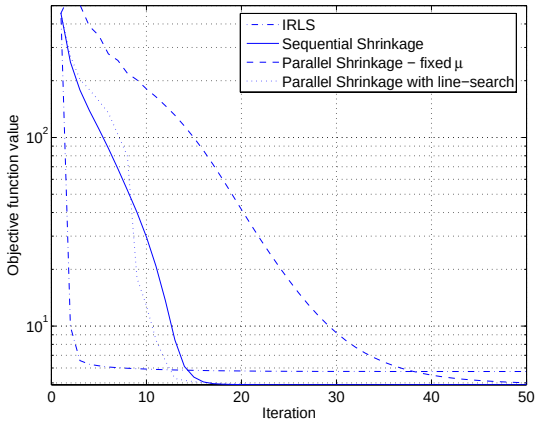

Figure 3: Experiment 1: The objective in Equation (10) as a function of the iteration – algorithms B-D.

1-st and the 5-th iterations) to the simple shrinkage *Algorithm A*. The simple shrinkage in this case uses a threshold being $\lambda/\alpha = 10\lambda$, based on Equation (30), so as to match to the objective function that uses λ in its formulation. Figure 4 (top-left) presents this comparison, showing the noise decay factor versus λ. Interestingly, it appears than the simple shrinkage manages to utilize most of the denoising potential, and 5 iterations of the IRLS give only slightly better results.

Figure 4 also presents similar comparisons of the simple shrinkage with the sequential coordinate descent (*Algorithm C*), and the parallel coordinate descent with line search or with a fixed μ chosen as $\mu = 1/\alpha$. First, we see that 5 iterations of the sequential CD are as effective as 5 IRLS iterations, giving better results than the simple shrinkage. Second, we see that the first iteration of the parallel shrinkage aligns perfectly with the simple shrinkage when $\mu = 1/\alpha$, as predicted, and having 5 iterations gives a slight improvement. Finally, as line search is introduced in *Algorithm D*, the results hardly change, implying that the choice $\mu = 1/\alpha$ is near-optimal.

Figure 5 presents the actual μ found by the line search in *Algorithm D* in the first iteration, as a function of the varying λ. We see that for small λ (where our assumptions in Section 3.6 hold true), the value found is close to $1/\alpha$ as expected.

Experiment 2 – A non-tight frame with normalized columns: We build $\mathbf{D}$ as a random matrix of size 100×1000 with entries drawn as Gaussian iid, and then normalize each column. The rest of the data generation follows the same procedure described for experiment 1.

Generally speaking, the results of this experiment are similar to those in Experiment 1, as Figure 6

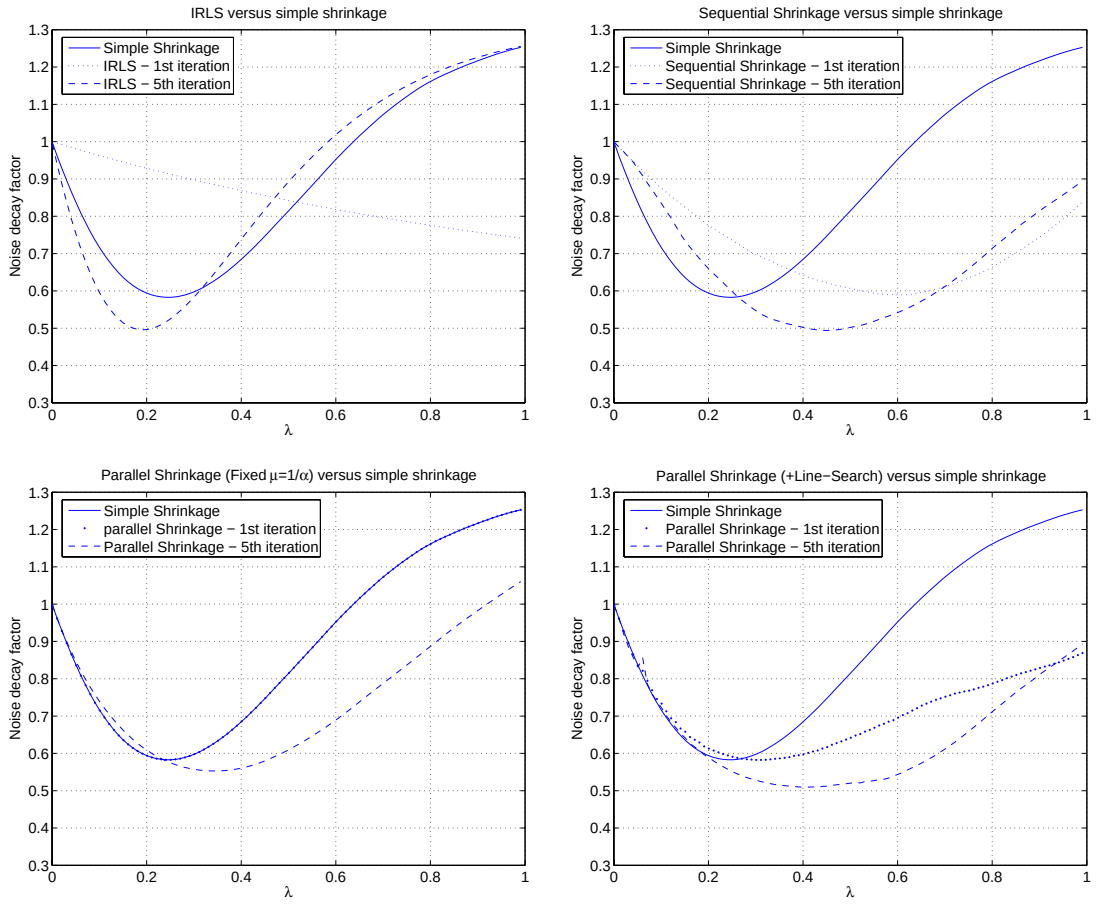

Figure 4: Experiment 1 results – comparing denoising effects of the various algorithms. Top left: the IRLS versus simple shrinkage; Top right: the sequential coordinate descent algorithm versus simple shrinkage; Bottom left: the parallel coordinate descent algorithm with fixed $\mu = 1/\alpha$ versus simple shrinkage; Bottom right: the parallel coordinate descent algorithm with line-search versus simple shrinkage.

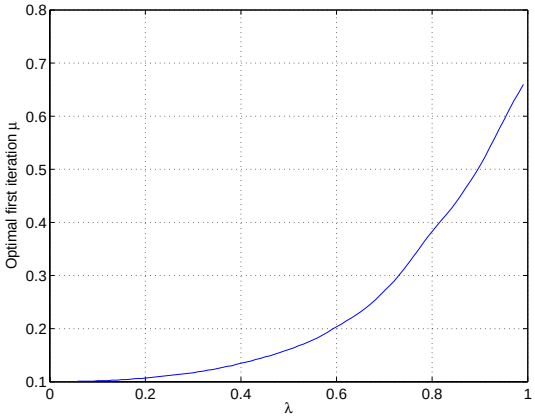

Figure 5: Experiment 1: The line-search results for μ in the first iteration.

suggests. Here we cannot align the choice of threshold in the simple shrinkage to the choice of λ in the objective function, simply because those two have not been related. We see that 5 iterations of the IRLS or the sequential CD can give a substantial improvement in denoising. Also, we see that the first iteration of *Algorithm D* that parallels in complexity to the simple shrinkage perform as good, and adding several iterations give further noise decay. Here we have not tried a fixed μ since the $1/\alpha$ rule does not apply.

Experiment 3 – A general frame: We build $\mathbf{D}$ as a random matrix of size 100×1000 with entries drawn as Gaussian iid. We deliberately change the scale of the columns to range linearly between 0.5 and 1. The rest of the data generation follows the same procedure described for experiment 1. Figure 7 presents how *Algorithm D* compares to the simple shrinkage. As can be seen again, while the alternative shrinkage formulation we get is different from the heuristic method, it has the same complexity and a comparable (and slightly better) noise decay. Performing 5 iterations further improves the performance slightly more.

Experiment 4 – Average performance: The results reported above correspond to one specific signal and the denoising obtained for it, so as to illustrate the relation between the various methods. We now introduce a wider experiment, where a corpus of signals is generated, contaminated by additive noise, and then denoised. Our objective here is to show the *average* amount of better denoising that can be expected when turning from the heuristic shrinkage (i.e., the first iteration of *algorithm D*) to several iterations of *algorithm D*.

Using the same dictionary as in Experiment 1, the signals in this experiment are generated by synthesizing a sparse representation $\mathbf{z}_0$ with L non-zeros, where $1 \leq L \leq 20$. Each such signal is normalized,

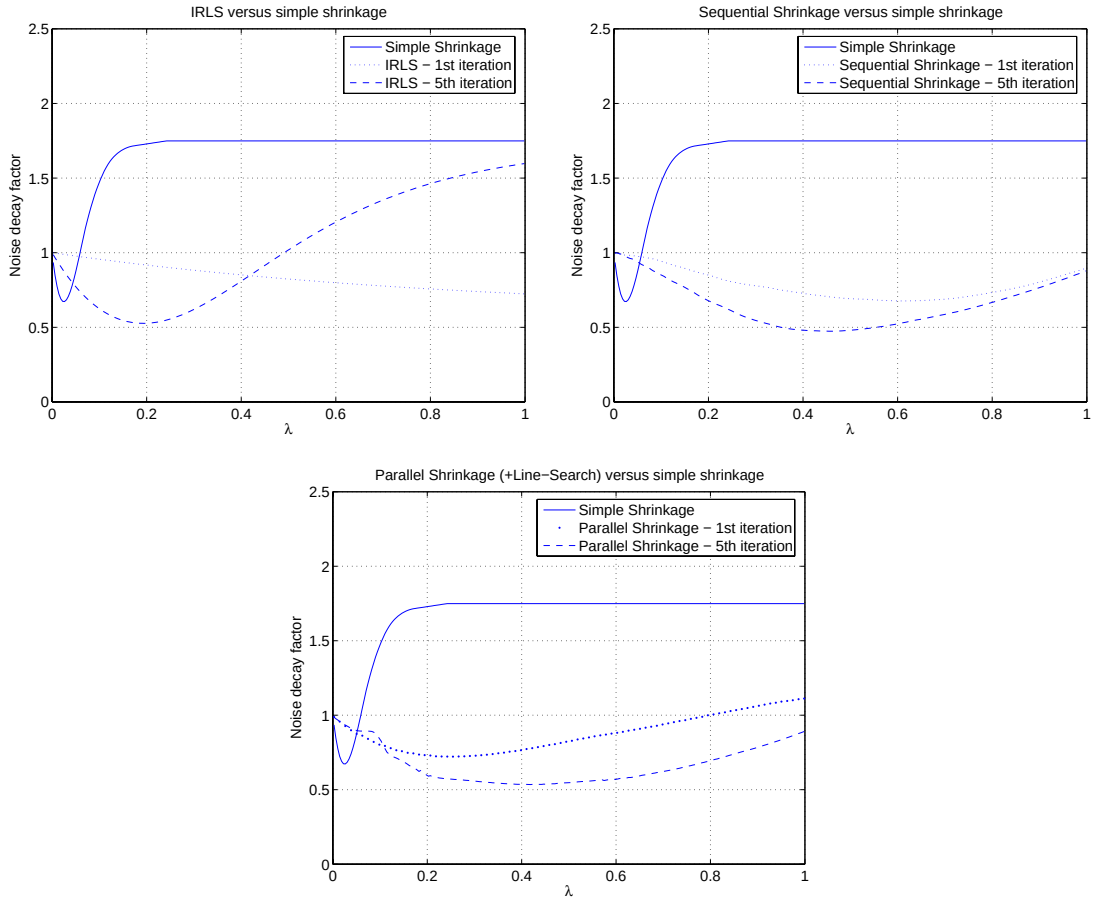

Figure 6: Experiment 2 results – comparing denoising effects of the various algorithms. Top left: the IRLS versus simple shrinkage; Top right: the sequential coordinate descent algorithm versus simple shrinkage; Bottom: the parallel coordinate descent algorithm with line-search versus simple shrinkage.

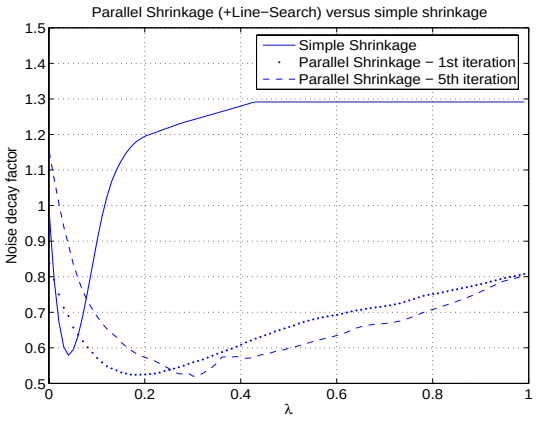

Figure 7: Experiment 3: The denoising effect of the parallel coordinate descent algorithm with line-search versus simple shrinkage.

and then contaminated by additive Gaussian noise with varying power in the range $\sigma = [0.03, 0.96]$ (i.e., signal to noise ratio in the range $[0, 30]$dB). Per each L and σ we generate 50 random signals, and apply denoising based on *Algorithm D* with $1 - 10$ iterations. Per each experiment we choose the optimal λ, as done in Experiment 1, so as to exclude its influence.

Figure 8 shows the average noise decay factors obtained per each L and input SNR. Several conclusions can be drawn from the results: (i) The denoising results are roughly the same for cardinalities in the range $[1, 20]$, implying that all these signals are sparse enough and thus handled similarly; (ii) The denoising effect depends on the SNR, showing better denoising results for a higher SNR; and (iii) Using more than 1 iteration in *Algorithm D* we typically get better performance with up to 2.5dB improvement. In some cases, more iterations may cause a deterioration in the denoising performance, but when this happens, it is a very mild loss (less than 0.05dB on average).

5 Related Work

Interestingly, a sequence of recent contributions proposed a similar sequential shrinkage algorithm. First, the work reported in [41, 42] uses such an algorithm for finding the sparsest representation over redundant dictionaries (such as the curvelet, or combination of dictionaries). These papers motivated such algorithm heuristically, relying on the resemblance to the unitary case, on one hand, and the block-coordinate-

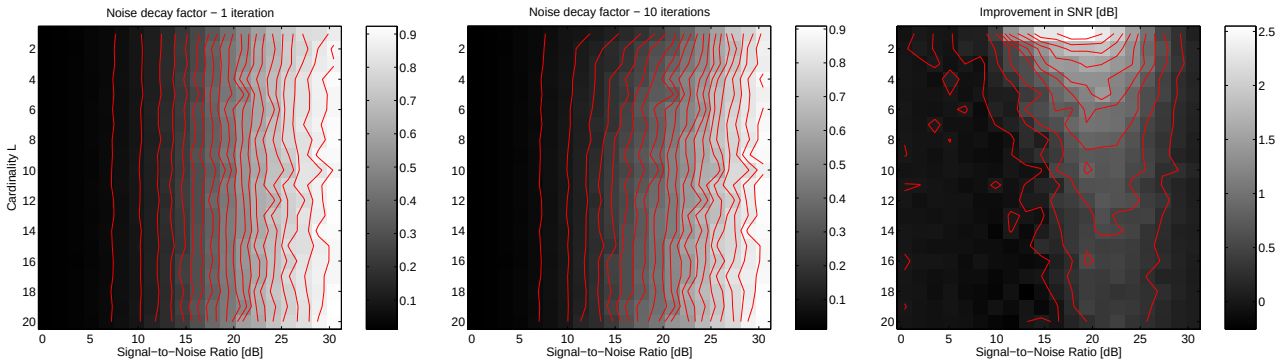

Figure 8: Experiment 4: Average denoising obtained by *Algorithm D* with 1 iteration (left), 10 iterations (middle), and their difference in dB (right). The results are shown as images with gray-values being proportional to the results. The color-bar on the right of each image shows the relation between brightness and resulting values. For convenience, the equi-height contours of the results are overlayed.

relaxation method, on the other [34].

Figueiredo and Nowak suggested a constructive method for image deblurring, based on iterated shrinkage [43]. Their algorithm aims at minimizing the penalty function

$$f_B(\mathbf{x}) = \frac{1}{2} \cdot \|\mathbf{K}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \cdot \mathbf{1}^T \cdot \rho\{\mathbf{W}\mathbf{x}\}, \tag{35}$$

where $\mathbf{K}$ is a (square) matrix representing the blur, and $\mathbf{W}$ is a unitary wavelet transform. Their sequential shrinkage method is derived via expectation-maximization, and its structure is very similar to the method proposed in this work. It turns out that their algorithm is not restricted to the case of square matrix $\mathbf{K}$, and as such can be generalized to handle the minimization of the objective posed in (9) by defining $\mathbf{D} = \mathbf{K}\mathbf{W}^H$.

Similarly, the paper by Daubechies, Defrise, and De-Mol [44] addresses the same objective as posed above, leaning on the definition of a sequence of surrogate functions, each minimized via shrinkage. This leads to yet another iterated shrinkage algorithm, very much like the one in [43].

While these two algorithms (EM-based and surrogate-based) are similar to ours, they are not the same. The norms of the atoms play different roles in these algorithms; the thresholds chosen in the shrinkage are somewhat different; and the choice of μ is done entirely different. Further work is required to clarify the relation between these methods.

6 Conclusion

In this paper we studied the heuristic shrinkage as is commonly practiced with redundant transforms. We have shown that such method has origins in Bayesian denoising, being the first iteration of an iterative denoising algorithm. This leads to several consequences: (i) we are now able to extend the heuristic shrinkage and get better denoising if more computations are allowed; (ii) we obtain alternative shrinkage algorithms that use the transform and its adjoint, rather than its pseudo-inverse; (iii) the new interpretation may help in addressing the question of choosing the threshold in shrinkage, and how to adapt it to the various coefficients, and (iv) the obtained algorithm can be used as an effective approximate solver for the BPDN for other applications, such as a non-linear transform that promotes sparsity.

We should emphasize that these findings are not to be confused as a recommendation to use shrinkage for denoising in its simple form. Treating each transform-coefficient alone is appealing because it is simple. However, recent work have shown that by treating clusters of coefficients, or exploiting the coefficients' inter-dependencies in other ways (e.g. Hidden-Markov models), could give a substantial improvement in the denoising effect [29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40].

Acknowledgements

The author would like to thank Dr. Michael Zibulevsky, Dr. Doron Shaked, and Prof. Yaacov Hel-Or for insightful discussions that helped in making this a better paper.

References

[1] Rudin, L., Osher, S., and Fatemi, C. (1992) Nonlinear total variation based noise removal algorithms, *Physica D*, Vol. 60, pp. 259–268.

[2] Carasso, A.S. (1999) Linear and nonlinear image deblurring: A documented study, *SIAM Journal On Numerical Analysis*, Vol. 36, pp. 1659–1689.

[3] Weickert, J., Haar Romeny, B.M., and Viergever, M.A. (1998) Efficient and reliable schemes for nonlinear diffusion filtering, *IEEE Transactions on Image Processing*, Vol. 7, No. 3, pp. 398–410, March.

[4] Elad, M. (2002) On the bilateral filter and ways to improve it, *IEEE Transactions on Image Processing*, Vol. 11, No. 10, pp. 1141–1151, October.

[5] Donoho, D.L and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika* Vol. 81 No. 3, pp. 425–455, September.

[6] Donoho, D.L. (1995) De-noising by soft thresholding, *IEEE Transactions on Information Theory*, Vol. 41, No. 3, pp. 613–627, May.

[7] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D (1995) Wavelet shrinkage - asymptopia, *Journal Of The Royal Statistical Society Series B - Methodological*, Vol. 57, No. 2, pp. 301–337.

[8] Donoho, D.L. and Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association*, Vol. 90 No. 432, pp. 1200–1224, December.

[9] Donoho, D.L. (1998) Wedgelets: Nearly minimax estimation of edges, *Annals Of Statistics*, Vol. 27, No. 3, pp. 859–897, June.

[10] Donoho, D.L. and Johnstone, I.M. (1998) Minimax estimation via wavelet shrinkage, *Annals of Statistics*, Vol. 26, No. 3, pp. 879–921, June.

[11] Simoncelli, E.P. and Adelson, E.H. (1996) Noise removal via Bayesian wavelet coring, Proceedings of the *International Conference on Image Processing*, Laussanne, Switzerland. September.

[12] Moulin, P. and Liu, J. (1999) Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors, *IEEE Transactions on Information Theory*, Vol. 45, No. 3, pp. 909-919, April.

[13] Jansen, M. (2001) *Noise Reduction by Wavelet Thresholding*, SpringerVerlag, New York.

[14] Candes, E.J. and Donoho, D.L. (2002) Recovering edges in ill-posed inverse problems: optimality of curvelet frames, *Annals of Statistics*, Vol. 30, No. 3, pp. 784–842, June.

[15] Starck, J.-L, Candes, E.J, and Donoho, D.L. (2002) The curvelet transform for image denoising, *IEEE Transactions On Image Processing*, Vol. 11, No. 6, pp. 670–684, June.

[16] Starck, J.-L., Elad, M., and Donoho D.L. (2004) Redundant multiscale transforms and their application for morphological component separation, *Advances in Imaging And Electron Physics*, Vol. 132, pp. 287–348.

[17] Do, M.N. and Vetterli, M. (2002) Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models, *IEEE Transactions On Multimedia* Vol. 4, No. 4, pp. 517–527, December.

[18] Do, M.N. and Vetterli, M. (2003) Framing pyramids, em IEEE Transactions On Signal Processing, Vol. 51, No. 9, pp. 2329–2342, September.

[19] Ishwar, P. and Moulin, P. (2000) Shift invariant restoration - an overcomplete maxent MAP framework, Proceedings of the *International Conference on Image Processing*, Vol. 3, pp. 270–272.

[20] Do, M.N. and Vetterli, M. (2003) The finite ridgelet transform for image representation, em IEEE Transactions On Image Processing, Vol. 12, No. 1, pp. 16–28 January.

[21] Carre, P. and Andres, E. (2004) Discrete analytical ridgelet transform, *Signal Processing*, Vol. 84, No. 11, pp. 2165–2173, November.

[22] Lang, M., Guo, H., and Odegard, J.E. (1996) Noise reduction using undecimated discrete wavelet transform, *IEEE Signal Processing Letters*, Vol. 3, No. 1, pp. 10–12, January.

[23] Chandrika, K., Fodor, I.K., and Gyaourova, A. (2002) Undecimated wavelet transforms for image denoising, *Lawrence Livermore National Laboratory*, technical report UCRL-ID-150931, November.

[24] Eslami, R. and Radha, H. (2003) The contourlet transform for image de-noising using cycle spinning, in Proceedings of *Asilomar Conference on Signals, Systems, and Computers*, pp. 1982–1986, November.

[25] Eslami, R. and Radha, H. (2004) Translation-invariant contourlet transform and its application to image denoising, submitted to the *IEEE Transactions on Image Processing*, October.

[26] Chen, S.S., Donoho, D.L. and Saunders, M.A. (2001) Atomic decomposition by basis pursuit, *SIAM Review*, Volume 43, number 1, pages 129–59.

[27] Karlovitz, L.A. (1970) Construction of nearest points in the ℓ^p, p even and ℓ^1 norms, *Journal of Approximation Theory*, Vol. 3, pp. 123-127.

[28] Gorodnitsky, I.F. and Rao, B.D. (1997) Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm, *IEEE Trans. On Signal Processing*, Vol. 45, No. 3, pp. 600-616, March.

[29] Scharcanski, J., Jung, C.R., and Clarke, R.T. (2002) Adaptive image denoising using scale and space consistency, *IEEE Transactions On Image Processing*, Vol. 11, No. 9, pp. 1092–1101, September.

[30] Jung, C.R. and Scharcanski, J. (2003) Adaptive image denoising and edge enhancement in scale-space using the wavelet transform, *Pattern Recognition Letters*, Vol. 24, No. 7, pp. 965–971, April.

[31] Portilla, J., Strela, V., Wainwright, M.J, and Simoncelli, E.P. (2001) Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain, Proceedings of the 8th *International Conference of Image Processing*, Thessaloniki, Greece. October.

[32] Portilla, J., Strela, V., Wainwright, M.J, and Simoncelli, E.P. (2003) Image denoising using scale mixtures of gaussians in the wavelet domain *IEEE Transactions On Image Processing*, Vol. 12, No. 11, pp. 1338–1351, November.

[33] Guleryuz, O.G. (2003) Weighted overcomplete denoising, Proceedings of the *Asilomar Conference on Signals and Systems*, Pacific Grove, CA, November.

[34] Bruce, A.G. , Sardy, S., and Tseng, P. (1998) Block coordinate relaxation methods for non-parametric signal de-noising. Proceedings of the SPIE - The International Society for Optical Engineering, Vol. 3391, pp. 75–86.

[35] Chang, S.G., Yu, B., and Vettereli, M. (2000) Adaptive wavelet thresholding for image denoising and compression, *IEEE Transactions on Image Processing*, Vol. 9, pp. 1532–1546, September.

[36] Chang, S.G., Yu, B., and Vettereli, M. (2000) Wavelet thresholding for multiple noisy image copies, *IEEE Transactions on Image Processing*, Vol. 9, pp. 1631–1635, September.

[37] Chang, S.G., Yu, B., and Vettereli, M. (2000) Spatially adaptive wavelet thresholding with context modeling for image denoising, *IEEE Transactions on Image Processing*, Vol. 9, pp. 1522–1530, September.

[38] Mrázek, P. and Weickert, J. (2003) Rotationally invariant wavelet shrinkage, in *Lecture Notes on Computer Science*, Vol. 2781, pp. 156–163.

[39] Steidl, G., Weickert, J., Brox, T., Mrázek, P., and Welk, M (2004) On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs, *SIAM Journal on Numerical Analysis*, Vol. 42, No. 2, 686–713.

[40] Weickert, J., Steidl, G., Mrázek, P., Welk, M., and Brox, T. (2005) Diffusion filters and wavelets: What can they learn from each other? In N. Paragios, Y. Chen, O. Faugeras (Eds.): *Mathematical Models of Computer Vision: The Handbook*.

[41] Starck, J.-L., Candes, E., and Donoho, D.L. (2003) Astronomical image representation by the curvelet transform, *Astronomy and Astrophysics*, Vol. 398, pp. 785–800.

[42] Starck, J.-L., Elad, M., and Donoho, D.L. (2004) Redundant multiscale transforms and their application for morphological component analysis, *Journal of Advances in Imaging and Electron Physics*, Vol. 132, pp. 287–348.

[43] Figueiredo, M.A. and Nowak, R.D. (2003) An EM algorithm for wavelet-based image restoration, *IEEE Trans. Image Process.* Vol. 12, No. 8, pp. 906–916, August.

[44] Daubechies, I., Defrise, M., and De-Mol, C. (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on Pure and Applied Mathematics*, Vol. LVII, pp. 1413-1457.

[45] Matalon, B., Elad, M. and Zibulevsky, M. (2005) Improved denoising of images using modeling of the redundant contourlet transform, Proceedings of the SPIE conference wavelets, Vol. 5914, July.

Biography

Michael Elad received his B.Sc, M.Sc. and D.Sc. from the department of Electrical engineering at the Technion, Israel, in 1986, 1988 and 1997 respectively. From 1988 to 1993 he served in the Israeli Air Force. From 1997 to 2000 he worked at Hewlett-Packard laboratories as an R&D engineer. From 2000 to 2001 he headed the research division at Jigami corporation, Israel. During the years 2001 to 2003 Michael was a research associate with the computer science department at Stanford university (SCCM program). Starting on September 2003, Michael is with the department of Computer science, the Technion, Israel Institute of Technology (IIT) as an assistant professor.

Michael Elad works in the field of signal and image processing, specializing in particular on inverse problems, sparse representations and over-complete transforms. Michael received the Technion's best lecturer award four times (1999, 2000, 2004, and on 2005). Michael is also the recipient of the Guttwirth and the Wolf fellowships. He is currently serving as an associate editor for IEEE Trans. on image processing, and EURASIP signal processing journals.