



HAL
open science

Thresholding RULES and iterative shrinkage/thresholding algorithm: A convergence study

Matthieu Kowalski

► **To cite this version:**

Matthieu Kowalski. Thresholding RULES and iterative shrinkage/thresholding algorithm: A convergence study . 21st IEEE International Conference on Image Processing (ICIP 2014), Oct 2014, La Défense, Paris, France. 10.1109/icip.2014.7025843 . hal-01102810

HAL Id: hal-01102810

<https://hal.science/hal-01102810>

Submitted on 13 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THRESHOLDING RULES AND ITERATIVE SHRINKAGE/THRESHOLDING ALGORITHM: A CONVERGENCE STUDY

Matthieu Kowalski

L2S, CNRS-SUPELEC-Univ Paris-Sud, Gif-sur-Yvette, France

ABSTRACT

Imaging inverse problems can be formulated as an optimization problem and solved thanks to algorithms such as *forward-backward* or *ISTA* (Iterative Shrinkage/Thresholding Algorithm) for which non smooth functionals with sparsity constraints can be minimized efficiently. However, the soft thresholding operator involved in this algorithm leads to a biased estimation of large coefficients. That is why a step allowing to reduce this bias is introduced in practice. Indeed, in the statistical community, a large variety of thresholding operators have been studied to avoid the biased estimation of large coefficients; for instance, the non negative Garrote or the the SCAD thresholding. One can associate a non convex penalty to these operators. We study the convergence properties of ISTA, possibly relaxed, with any thresholding rule and show that they correspond to a semi-convex penalty. The effectiveness of this approach is illustrated on image inverse problems.

Index Terms— Sparse approximation, semi convex optimization, nonnegative garrote, relaxed ISTA

1. INTRODUCTION: MATHEMATICAL FRAMEWORK AND STATE OF THE ART

Wavelet thresholding Before the expansion of the sparse principle in signal and image processing and statistics, wavelet thresholding [1] has been intensively studied and has provided nice results for image denoising [2]. The survey of A. Antoniadis on wavelet decomposition in [3] rigorously defines the notion of the “*thresholding rule*” and presents various thresholding operators. A major results of [3], is that one can explicitly compute a (non necessarily unique) penalty to a given thresholding rule.

The two most popular thresholding rules are probably the Soft and the Hard Thresholding. However, these two rules suffers from two different shortcomings: the Hard thresholding induces a discontinuity, but preserves an unbiased estimation of large coefficients when the Soft thresholding is continuous but induces a biased estimation of large coefficients. To circumvent these drawbacks, various thresholding rules were introduced, such as Non Negative Garotte, SCAD, Firm thresholding and many others (see [3]).

Iterative Shrinkage/Thresholding Algorithm Wavelet thresholding is well defined in the context of orthogonal transformations, but its “natural” extension to inverse problems or redundant transformations is done thanks to the Basis-Pursuit/Lasso [4, 5]. This connection was nicely stressed in [6] with the following statement: “simple shrinkage could be interpreted as the first iteration of an

algorithm that solves the basis pursuit denoising (BPDN) problem”. Indeed, it is now widely known that ℓ_1 regularized problems can be minimized thanks to the forward-backward [7], or ISTA [8], algorithm.

By extension, this algorithm – restated in Alg.1, can handle the following convex problem, where $\mathbb{S}_P(\cdot; \lambda)$ stands for the proximity operator of λP^1 :

$$\min_{x \in \mathbb{R}^N} f(x) + \lambda P(x) \quad (1)$$

where

- $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is a proper convex lower semi-continuous function, L -Lipschitz differentiable, called the “Loss”.
- $P : \mathbb{R}^N \rightarrow \mathbb{R}$ is a non-smooth proper convex lower semi-continuous function, called the “Penalty”.
- $f + \lambda P, \lambda \geq 0$ is a coercive finite function.

An appropriate choice of the relaxation parameters $\gamma^{(k)}$ ensures fast convergence to a minimum of the functional in $\mathcal{O}(1/k^2)$ iterations. For a constant choice $0 \leq \gamma < 1$, the convergence toward a minimum is guaranteed in $\mathcal{O}(1/k)$ iterations.

Algorithm 1: relaxed ISTA

Input: $x^{(0)}, z^{(0)} \in \mathbb{R}^N, \mu < \frac{1}{L}$

repeat

$$\left| \begin{array}{l} x^{(k+1)} = \mathbb{S}_{\mu P} \left(z^{(k)} - \mu \nabla f(z^{(k)}); \lambda \right); \\ z^{(k+1)} = x^{(k+1)} + \gamma^{(k)} (x^{(k+1)} - x^{(k)}); \end{array} \right.$$

until convergence;

However, if the ℓ_1 minimization takes benefit from the convex optimization framework, the problem of biased estimation of large coefficients remains. That is why in practice, one can use a “debiasing step” [9]. The Iterative Hard-Thresholding [10] can also reduce this bias.

In [11], the authors study the convergence of a general forward-backward scheme, under the assumption that the function satisfies the Kurdyka-Lojasiewicz inequality. While the authors claim that this condition is satisfied by a wide range of functions, it can be difficult to check it in practice. In [12], the convergence of ISTA (without relaxation) in a non-convex setting is performed for an ℓ_2 loss, using a classical bounded curvature condition (BBC) on the penalty term. It is shown that this condition is verified by the penalty associated to some classical thresholding rules, such as Hard/Soft-thresholding or SCAD thresholding. The author extends his results to group-Thresholding with a loss belonging to the natural exponential family in [13]. In [14], the authors study the ISTA algorithm with a line search, when f is L -Lipschitz differentiable (possibly non convex) and P can be written as a difference of two convex functions.

$${}^1\mathbb{S}_P(x; \lambda) = \text{prox}_{\lambda P}(x) = \underset{\alpha \in \mathbb{R}^N}{\text{argmin}} \frac{1}{2} \|x - \alpha\|^2 + \lambda P(\alpha)$$

MK benefited from the support of the “FMJH Program Gaspard Monge in optimization and operation research”, and from the support to this program from EDF.

Contributions and outline In this article, we widely extend the results of [12, 13] by studying the convergence of Alg. 1 with any thresholding rules, for any L -Lipschitz differentiable Loss, with a possible relaxation parameter. The link with non convex optimization that we establish thanks to the notion of semi-convex functions allows to provide a convergence result (Theorem 5) on ISTA, where the hypotheses are made directly on the thresholding rule instead of the functional to optimize. Section 2 is a reminder of some important definitions and establishes some important properties of thresholding rule and semi-convex function. As it can be convenient in practice to use a constant step size in the algorithm, as well as to relax it, we cannot use directly the results given in [14]. Main convergence results are proven in Section 3. Finally, Section 4 provides numerical experiments on image deconvolution and inpainting.

2. THRESHOLDING RULE AND SEMI-CONVEXITY

2.1. Semi-convex functions

An important tool for proving the convergence results is the notion of semi-convex function (see for example [15] and references therein), which is reminded here.

Definition 1 (c -semi-convex functions). *A function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is said c -semi-convex if there exists a finite constant $c \geq 0$ such that*

$$\tilde{f}(x) := f(x) + \frac{c}{2}\|x\|^2$$

defines a convex function.

Remark 1. *When $c = 0$, f is actually a convex function. The case $c < 0$ could be included in the definition, f being then c -strongly convex.*

One can remark that on \mathbb{R}^N , semi-convex functions are continuous. We can define the subdifferential of a semi-convex function, as for convex functions, which is simply given thanks to the subdifferential of \tilde{f} by

$$\partial_f(x) = \partial_{\tilde{f}}(x) - cx.$$

We also extend the notion of proximity operator for f by:

$$\text{prox}_f(y) = \underset{x \in \mathbb{R}^N}{\text{argmin}} \frac{1}{2}\|y - x\|^2 + f(x).$$

In [12], the main hypothesis to ensure the convergence of ISTA relies on the fact that the penalty term satisfies the BCC:

Definition 2 (Bounded curvature condition). *A function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ verifies the bounded curvature condition (BCC) if there exists a symmetric matrix H such that*

$$f(x + y) \geq f(x) + \langle y, s \rangle - \frac{1}{2}y^T H y, \quad \forall y \in \mathbb{R}^N$$

where $s = z - \text{prox}(z)$, with z is such that $x = \text{prox}(z)$.

For semi-convex functions, we have that $s \in \partial_f(x)$ in the previous definition. Next proposition links semi-convexity and the BCC.

Proposition 1. *Semi-convexity is a sufficient condition for BCC to hold.*

Proof. Let $x \in \mathbb{R}^N$, and f be a c -semi-convex function. Let \tilde{f} defined as in Def. 1. Then, for any $s \in \partial_{\tilde{f}}(x)$ we have $\tilde{f}(x + h) - \tilde{f}(x) \geq \langle s, h \rangle$. Then

$$f(x + h) - f(x) \geq \langle s - cx, h \rangle - \frac{c}{2}\|h\|^2$$

with $s - cx \in \partial_f(x)$ so that f satisfies the BCC. ■

2.2. Thresholding rules

We first recall the definition of a thresholding rule that allows to establish the link with a semi-convex penalty.

Definition 3 (Thresholding rule). $\mathbb{S}(\cdot; \lambda)$ is a thresholding rule iff

1. $\mathbb{S}(\cdot; \lambda)$ is an odd function. $\mathbb{S}_+(\cdot; \lambda)$ denotes its restriction to \mathbb{R}_+
2. $0 \leq \mathbb{S}_+(x; \lambda) \leq x, \forall x \in \mathbb{R}_+$
3. \mathbb{S}_+ is nondecreasing on \mathbb{R}_+ , and $\lim_{x \rightarrow +\infty} \mathbb{S}_+(x; \lambda) = +\infty$

Following [3], we can associate a penalty P to any thresholding rule, such that $\mathbb{S} = \text{prox}_P$ with

$$P(x; \lambda) = \int_0^{|x|} \mathbb{S}_+^{-1}(t; \lambda) - t \, dt.$$

where $\mathbb{S}^{-1}(t; \lambda) = \sup\{x; \mathbb{S}(x; \lambda) \leq t\}$ and $\mathbb{S}^{-1}(-t; \lambda) = -\mathbb{S}^{-1}(t; \lambda)$. We can now show that such a penalty is semi-convex thanks to the following theorem.

Theorem 1. *The penalty $P(\cdot; \lambda)$ associated to a thresholding rule $\mathbb{S}(\cdot; \lambda)$ is at least 1 -semi-convex. Moreover, if \mathbb{S} is continuous and the difference quotient is bounded by a , then P is at least $1 - \frac{1}{a}$ semi-convex. In addition, P is ℓ -strongly convex iff $a \leq \frac{1}{1+\ell}$.*

Proof. P is differentiable almost everywhere, with derivative (for $x > 0$): $P'(x; \lambda) = \mathbb{S}_+^{-1}(x) - x$. Then, one can check that $x \mapsto P(x; \lambda) + \frac{1}{2}\|x\|^2$ admits a right derivative which is nondecreasing. Then, P is 1 -semi-convex.

Suppose \mathbb{S} is continuous with a difference quotient bounded by a . Let $0 < t_1 < t_2$, we have that $\mathbb{S}_+^{-1}(t_1) \leq \mathbb{S}_+^{-1}(t_2) - \frac{t_2 - t_1}{a}$, then $x \mapsto P(x; \lambda) + (1 - \frac{1}{a})\|x\|^2$ is convex, hence the conclusion. ■

2.3. Examples of thresholding rules

We give here three well-known examples of thresholding rules: Hard, SCAD and NonNegative Garotte (NNG) thresholdings, with their associated Penalties. These thresholding rules are plotted on Fig. 2.3. One can see that the SCAD and NNG thresholdings can be viewed as two different kinds of ‘‘compromise’’ between the Soft and Hard thresholdings.

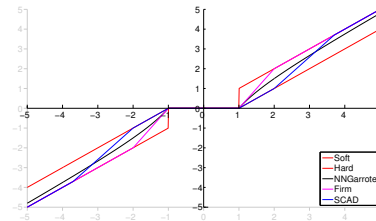


Fig. 1. Some thresholding rules

Hard-thresholding: $\mathbb{S}(x; \lambda) = \begin{cases} 0 & \text{if } |x| \leq \lambda \\ x & \text{if } |x| > \lambda \end{cases}$

$$P'(x, \lambda) = \begin{cases} -x + \lambda \text{sgn}(x) & \text{if } |x| \leq \lambda \\ x & \text{if } |x| > \lambda \end{cases} \quad P(x, \lambda) = \begin{cases} -\frac{x^2}{2} + \lambda|x| & \text{if } |x| \leq \lambda \\ \frac{\lambda^2}{2} & \text{if } |x| > \lambda \end{cases}$$

As P is not continuous, P is 1 -semi-convex.

$$\text{SCAD: } \mathbb{S}(x; \lambda) = \begin{cases} x \left(1 - \frac{\lambda}{|x|}\right)^+ & \text{if } |x| \leq 2\lambda \\ \frac{x}{a-2} \left(a - 1 - \frac{a\lambda}{|x|}\right) & \text{if } 2\lambda < |x| \leq a\lambda \\ 0 & \text{if } |x| > a\lambda \end{cases}$$

with $a > 2$.

$$P'(x) = \begin{cases} \lambda & \text{if } x \leq \lambda \\ \frac{a\lambda - x}{a-1} & \text{if } \lambda < x \leq a\lambda \\ 0 & \text{if } x > a\lambda \end{cases} \quad P(x; \lambda) = \begin{cases} \lambda x & \text{if } x \leq \lambda \\ \frac{(a\lambda x - x^2/2)}{a-1} & \text{if } \lambda < x \leq a\lambda \\ a\lambda & \text{if } x > a\lambda \end{cases}$$

Looking at P' , the SCAD penalty is semi-convex with $c = \frac{1}{a-1}$.

$$\text{Nonnegative Garrote: } \mathbb{S}(x; \lambda) = x \left(1 - \frac{\lambda^2}{x^2}\right)^+$$

$$P'(x) = \frac{2\lambda^2}{\sqrt{x^2 + 4\lambda^2} + |x|} \text{sgn}(x)$$

$$P(x; \lambda) = \lambda^2 + \text{asinh}\left(\frac{|x|}{2\lambda}\right) + \lambda^2 \frac{|x|}{\sqrt{x^2 + 4\lambda^2} + |x|}$$

Looking at P' , the NNG penalty is semi-convex with $c = \frac{1}{2}$.

3. ISTA WITH ANY THRESHOLDING RULE

We are now ready to show the convergence of ISTA when \mathbb{S} is not necessarily a proximity operator anymore but a thresholding rules as defined in Def. 3. From an optimization point of view, one can consider the problem of minimizing

$$F(x) = f(x) + P(x; \lambda) \quad (2)$$

with f being a L -Lipschitz differentiable function and P a c -semi-convex function. Then, one can state

Theorem 2. *If $\mu < \frac{2}{L+c}$, then any convergent subsequences of $\{x^{(k)}\}$ generated by Alg. 1 ISTA, with $\gamma^{(k)} = 0$ for all k , converges to a critical point of F .*

Proof. The proof classically relies on the global convergent theorem [16]. We only prove that we have a ‘‘descent function’’, the two other points being straightforward thanks to the continuity of the functional.

$$\text{Let } \mathcal{M}_y(x) = \frac{1}{2}\|y-x\|^2 + \mu P(x; \lambda) \text{ and } x^* = \underset{x}{\text{argmin}} \mathcal{M}_y(x).$$

Then, one has thanks to the semi-convexity of P

$$\begin{aligned} \mathcal{M}_y(x^* + h) - \mathcal{M}_y(x^*) &= \\ \frac{1}{2}\|y - x^* - h\|^2 + \mu P(x^* + h; \lambda) - \frac{1}{2}\|y - x^*\|^2 - \mu P(x^*; \lambda) &= \\ \frac{1}{2}\|h\|^2 - \langle y - x^*, h \rangle + \mu P(x^* + h; \lambda) - \mu P(x^*; \lambda) &\geq \\ \geq \frac{1 - \mu c}{2} \|h\|^2 \end{aligned}$$

Let $\ell_F(x; y) = f(y) + \langle \nabla f(y), x - y \rangle + P(x; \lambda)$, then

$$\begin{aligned} x^{(k+1)} &= \underset{x}{\text{argmin}} \frac{1}{2\mu} \|x^k - x\|^2 + \ell_F(x; x^k) \\ &= \underset{x}{\text{argmin}} \frac{1}{2} \|x^{(k)} - \mu \nabla f(x^{(k)}) - x\|^2 + \mu P(x; \lambda) \end{aligned}$$

and

$$\ell_F(x^{(k+1)} + h; x^k) - \ell_F(x^{(k+1)}; x^k) \geq -\frac{c}{2} \|h\|^2 + \langle x^{(k)} - x^{(k+1)}, h \rangle \quad (3)$$

As f is L -Lipschitz differentiable, we have

$$F(x^{(k+1)}) \leq \ell_F(x^{(k+1)}; x^{(k)}) + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2. \quad (4)$$

Then, using Eq.(3) and (4), with $h = x^{(k)} - x^{(k+1)}$ we have

$$F(x^{(k+1)}) \leq F(x^{(k)}) - \frac{2/\mu - c - L}{2} \|x^{(k+1)} - x^{(k)}\|^2$$

Then, as soon as $2/\mu > L + c$, one can apply the global convergence theorem, hence the conclusion. ■

Theorem 3. *Suppose f is ℓ -strongly convex. Then, if $c \leq \ell$, any accumulation point of $\{x^{(k)}\}$ is a global minimizer of F .*

Proof. If f is ℓ -strongly convex and P is c -semi-convex, then F is $c - \ell$ -semi-convex. Then, as soon as $\ell \geq c$, F is convex, hence the conclusion. ■

Theorem 4. *Let the step size $\mu = 1/L$. If f is convex and $\gamma < \sqrt{1 - c/L}$, then any convergent sub-sequence of $\{x^{(k)}\}$ generated by relaxed ISTA converges to a critical point of F .*

Proof. Let $z^{(k)} = x^{(k)} + \gamma(x^{(k)} - x^{(k-1)})$ with

$$x^{(k+1)} = \underset{x}{\text{argmin}} \frac{1}{2L} \|z^k - x\|^2 + \ell_F(x; z^k)$$

we have

$$\begin{aligned} F(x^{(k+1)}) &\leq \ell_F(x^{(k+1)} + h; z^{(k)}) + \frac{L}{2} \|x^{(k+1)} - z^{(k)}\|^2 + \\ \frac{c}{2} \|h\|^2 - L \langle z^{(k)} - x^{(k+1)}, h \rangle. \end{aligned}$$

Wich, with $h = x^{(k)} - x^{(k+1)}$, gives

$$F(x^{(k+1)}) + \frac{L-c}{2} \|x^{(k+1)} - x^{(k)}\|^2 \leq \ell_F(x^{(k)}; z^{(k)}) + \frac{L\gamma^2}{2} \|x^{(k)} - x^{(k-1)}\|^2.$$

If f is convex, then $\ell_F(x^{(k)}; z^{(k)}) \leq F(x^{(k)})$, then, if $\gamma < \sqrt{1 - c/L}$ one can apply the global convergence theorem [16], hence the conclusion. ■

However, it is much more easier to choose the appropriate thresholding rules instead of its associated non-convex penalty. As a corollary of the previous Theorems, one can state

Theorem 5. *Let \mathbb{S} being a thresholding rule, and f a convex L -Lipschitz differentiable function. Then ISTA, ie. Alg.1 with $\gamma = 0$, converges. Moreover, if $\mathbb{S}' \leq a$, then relaxed ISTA, ie Alg.1, with $\gamma < \frac{1}{a}$ converges.*

Remark 2. *With the hard-thresholding rule, one must choose the step size $\mu < 1/L$, and the convergence result on the relaxed algorithm does not apply.*

4. NUMERICAL ILLUSTRATION

We provide here two numerical illustrations on image restoration (deblurring + denoising) and inpainting. The aim of this section is not to demonstrate that the choice of a thresholding operator is better than others, but to show that soft and hard thresholding are not the only possible choices.

The two problems can be formulated as a linear inverse problem. Using a sparse synthesis model in a undecimated wavelet frame, it can be formulated as:

$$y = \Omega s_0 + n = \Omega \Psi \alpha + n$$

where s_0 is the original image, Ω is linear operator corresponding to the blurring kernel (for the image restoration problem) or the mask of random pixel locations (for inpainting problem). Ψ corresponds to a linear undecimated wavelet transform and α are synthesis coefficients of s_0 . n is a white gaussian noise. The chosen image for the two experiments is a part of the boat image of size 256×256 .

For the two experiments, ISTA was run with various thresholding operators, with 10 decreasing values of λ on a logarithmic scale between $\|\Psi^* \Omega^* y\|_\infty / L$ and $10^{-2} / L$, with $L = \|\Psi\|^2$. The choice of $\nabla f(\alpha)$ is the canonical choice $-\Psi^* \Omega (y - \Omega^* \Psi \alpha)$ corresponding to a ℓ_2 data term between the observed image y , and the synthesis coefficients α . 500 iterations of ISTA are run for each λ , with warm start. The relaxation parameter is 0.9 for Soft-Thresholding, 0.49 for NNG and 0 for Hard-Thresholding.

4.1. Image restoration

Here Ω corresponds to a Gaussian blur kernel with variance 1.2. n is a white Gaussian noise with variance 0.1. The resulting output SNR is 14.90. The restored images are displayed on Fig. 2, λ is such that the empirical variance of the residual is close to 0.1 (known as the Morozov's discrepancy principle [17]).

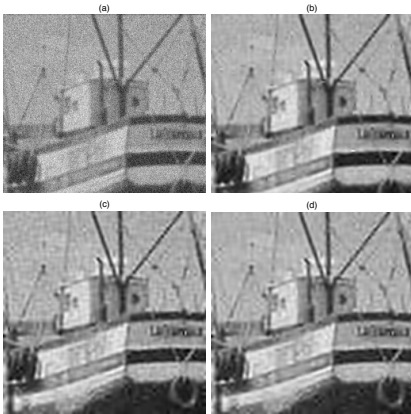


Fig. 2. (a) degraded image (b) Soft (SNR = 20 dB). (c) NNG (SNR = 19.3 dB). (d) Hard (SNR = 19.6 dB).

On these experiments, one can see that the performances of all the thresholding operators are sensibly equivalent with a slight advantage for the Soft-Thresholding. SCAD and Firm obtain similar results (not shown here).

4.2. Image inpainting

Here Ω corresponds to a mask of random pixel locations. The resulting image has 90% missing pixels, with no additional noise. The restored images are displayed on Fig. 3, for $\lambda \rightarrow 0$. The given SNR are computed on the missing pixels only.

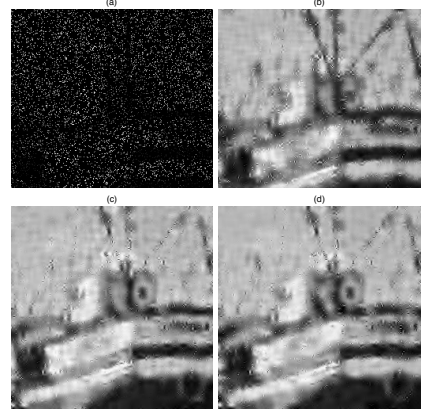


Fig. 3. (a) Degraded Image (b) Soft (SNR = 12.9 dB). (c) NNG (SNR = 14.6 dB). (d) Hard (SNR = 14.3 dB).

For this problem, the performances are inverted compared to the previous one: NNG and Hard-Thresholding give the bests SNR, while Soft gives the less satisfactory reconstruction. SCAD and Firm obtain similar results than NNG (not shown here).

5. DISCUSSION AND CONCLUSION

The results obtained to the previous toy problem confirm that good performances can be achieved with various thresholding rules, and the choice between Soft or Hard thresholding depends of the considered problem. The choice of NNG (or SCAD, or Firm) thresholding rules appears to be a good compromise between this two operators. However, the results are not as impressive as in [18], where NNG was heuristically used inside ISTA for declipping audio signal, or in [18] where NNG was compared to other thresholding rules on audio denoising. In these two papers, NNG outperforms both Hard and Soft Thresholdings.

Still, the main objective of this article is to provide theoretical guarantee for using (relaxed)-ISTA with thresholding rules. One can notice that in [19], it was shown that an algorithm similar to relaxed-ISTA can be used in the context of nonconvex optimization. The main difference being that the algorithm studied in [19] involves several hyper-parameters when Alg. 1 has no parameter except the relaxation parameter.

The study made here can be extended straightforward to independent group shrinkage, following [13] which extends the approach of Antoniadis [3]. Generalization to data term f which is not Lipschitz differentiable seems also possible, using an adaptive step size instead of the constant step size equal to $1/L$, following the analysis made by Tseng and Yun for block-coordinate descent in [20].

Finally, convergence analysis of ISTA with “thresholding” operators with overlap, as the windowed-group-Lasso [21] remains an open problem.

6. REFERENCES

- [1] D. L. Donoho, "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data," in *Proceedings of symposia in Applied Mathematics*, vol. 47. Providence: American Mathematical Society, 1993, pp. 173–205.
- [2] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *Image Processing, IEEE Transactions on*, vol. 9, no. 9, pp. 1532–1546, 2000.
- [3] A. Antoniadis, "Wavelet methods in statistics: Some recent developments and their applications," *Statistics Surveys*, vol. 1, pp. 16–55, 2007.
- [4] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Serie B*, vol. 58, no. 1, pp. 267–288, 1996.
- [6] M. Elad., "Why simple shrinkage is still relevant for redundant representations?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5559–5569, 2006.
- [7] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, Nov. 2005.
- [8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [9] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 586–597, 2007.
- [10] M. E. D. T. Blumensath, "Iterative thresholding for sparse approximations," *The Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654, 2008.
- [11] J. Bolte, H. Attouch, and B. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel method," *Mathematical Programming*, vol. 137, no. 1–2, pp. 91–129, 2013.
- [12] Y. She, "Thresholding-based iterative selection procedures for model selection and shrinkage," *Electronic Journal of Statistics*, vol. 3, 2009.
- [13] ———, "An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors," *Computational Statistics & Data Analysis*, vol. 56, no. 10, pp. 2976–2990, 2012.
- [14] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," *Journal of Machine Learning Research*, vol. 28, no. 2, pp. 37–45, 2013.
- [15] A. Colesanti and D. Hug, "Hessian measures of semi-convex functions and applications to support measures of convex bodies," *manuscripta mathematica*, vol. 101, no. 2, pp. 209–238, 2000.
- [16] M. Bazaraa and C. Shetty, *Nonlinear Programming: Theory and Algorithms*, Wiley ed., New York, 1979.
- [17] V. A. Morozov, "On the solution of functional equations by the method of regularization," *Soviet Math. Dokl.*, vol. 7, pp. 414–417, 1966.
- [18] K. Siedenburg, M. Dörfler, and M. Kowalski, "Audio declipping with social sparsity," *Preprint*, 2013.
- [19] P. Ochs, Y. Chen, T. Brox, and T. Pock, "ipiano: Inertial proximal algorithm for non-convex optimization," *Preprint*, 2013.
- [20] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Math. Prog. B*, vol. 117, pp. 387–423, 2009.
- [21] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE transactions on signal processing*, vol. 61, no. 10, pp. 2498–2511, 2013.