Robustifying stability of the Fast iterative shrinkage thresholding algorithm for ℓ_1 regularized problems

Gustavo Silva
Electrical and Computer Engineering
Pontificia Universidad Católica del Perú
Lima, Peru
Email: gustavo.silva@pucp.edu.pe

Paul Rodriguez
Electrical and Computer Engineering
Pontificia Universidad Católica del Perú
Lima, Peru
Email: prodrig@pucp.edu.pe

***Abstract*—The fast iterative shrinkage-thresholding algorithm (FISTA) is a well-known first order method used to minimize ℓ_1 regularized problems. However, it is also a non-monotone algorithm that can exhibit a sudden and gradual oscillatory behavior during the convergence. One of the parameters that directly affects the convergence of the FISTA method, whose optimal value is typically unknown, is the step-size (SS) that is linked to the Lipschitz constant. Depending on a suitable selection of the SS either manual or automatic, and the SS evolution throughout iterations, e.g. constant, decreasing, or increasing sequence, the practical performance can differ in orders of magnitude with or without stability issues (oscillations or, in the worst case, divergence).**

In this paper, we propose an algorithm, which has two variants, to address the stability issues in case of ill-chosen parameters for a given SS policy (either manual or adaptive). The proposed method structurally consists of an instability prediction rule based on the ℓ_∞ norm of the gradient, and a correction for it, which can interpreted as an under-relaxation technique.

***Index Terms*—FISTA, stable convergence, step-size, convolutional sparse representation**

I. Introduction

Sparse models such as the ℓ_1 regularization [1] have consistently received increasing attention in signal processing, image processing and machine learning. This sparse models can be raised as a convex optimization problem of the form

$$\min_{\mathbf{x}\in\mathbb{R}^N} F(\mathbf{x}) := f(\mathbf{x}) + \lambda \cdot h(\mathbf{x}), \tag{1}$$

where $f, h : \mathbb{R}^N \mapsto \mathbb{R}$ are both convex functions, ∇f is L-Lipschitz continuous and h is such that it induces sparsity and has a computationally simple proximal operator, i.e. $\text{prox}_{\lambda h}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + \lambda \cdot h(\mathbf{x})$. While there are many methods to solve (1), e.g. ADMM [2], FISTA [3], IRLS [4], forward-backward splitting [5] and others, FISTA is widely used due to its simplicity and applicability in large-scale problems, generating the iterative sequence

$$\mathbf{x}_k = \text{prox}_{\alpha_k h}(\mathbf{y}_k - \alpha_k \nabla f(\mathbf{y}_k)) \tag{2}$$

$$\mathbf{y}_k = \mathbf{x}_k + \gamma_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \tag{3}$$

where $\alpha_k \in (0, \frac{1}{L}]$ is the step-size (SS) sequence, and $\gamma_k = \frac{t_k - 1}{t_{k+1}}$, referred to as the inertial sequence, is a weighting parameter that satisfies: $t_{k+1}^2 - t_{k+1} \leq t_k^2 \;\; \forall k \geq 1$.

Unfortunately, FISTA is a non-monotone algorithm that can exhibit oscillations during convergence [6], [7] of the objective. Since its rate of convergence (RoC) is directly related to the SS (see Section II-A), a poor estimation or selection of this parameter can cause stability issues as oscillations and, in the worst case, divergence. Compared to a direct objective function monitoring, [8] has recently been shown that gradient monitoring can help to anticipate this kind of erratic behavior.

In this paper, we introduce a robustifying stability algorithm, which has two variants, to effectively prevent potential stability issues. The proposed algorithm is assessed in the convolutional sparse coding and convolutional dictionary learning problems, which are particular cases of ℓ_1 regularized problems.

II. Preliminaries

A. FISTA-F3K: an improved FISTA variant

FISTA is one particular variant among several accelerated methods to solve problem (1), which theoretical RoC is inversely proportional α_k and t_k^2, the SS and the inertial sequence respectively, i.e.

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \varsigma \cdot \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{\alpha_k t_k^2}, \tag{4}$$

for some constant ς. In particular, the well-known $\mathcal{O}(k^{-2})$ FISTA's RoC, i.e. $F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \varsigma \cdot \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{k^2}$, is achieved (see [3, Section 4]) by selecting an inertial sequence such $t_k \geq \frac{k+1}{2}$, and by considering $\alpha_k \geq \alpha_{k+1}$. Past attempts to improve FISTA's RoC include alternative inertial sequences [9], [10], and intertwining the selection of the inertial sequence and the SS [11], [12]. Very recently, [13] heuristically showed that FISTA can achieve a RoC proportional to k^{-3} for the indexes where the SS exhibits an approximate linear growth, with the default $\mathcal{O}(k^{-2})$ behavior when the SS's bound is reached. In order to generate such exceptional step-size sequence, [13] proposed to use a modified version of the Cauchy SS

$$\alpha_k = c \cdot \frac{\|\mathbf{s}_k \odot \mathbf{g}_k\|_2^2}{\|\mathcal{A}(\mathbf{s}_k \odot \mathbf{g}_k)\|_2^2} \tag{5}$$

where $c \in (0, 1]$ is an small multiplicative factor, $\mathbf{s}_k = \text{supp}(\mathbf{x}_k) = I_{[|\mathbf{x}_k|>0]}$ is the support of the current (sparse) solution, $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$ is the gradient of f, and $\mathcal{A}$ is the forward operator associated to $f(\mathbf{x}) = 0.5 \cdot \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|_2^2$.

While (5) certainly has better convergence performance when compared to other adaptive schemes such Cauchy [14, Section 3] and Barzilai-Borwein [15], the factor c must be correctly chosen otherwise the resulting SS can force the objective to have an erratic behavior.

B. Gradient's behavior in descent methods

Although the classic target in convex optimization is to analyze the RoC of the functional value in (1), i.e. $F(\mathbf{x}_k)$, towards its optimal solution $F(\mathbf{x}^*)$, examining the evolution of $\nabla f(\mathbf{x}_k)$ [16] has also attracted interest in order provide additional strategies to improve practical performance of gradient based methods.

In the case of the (accelerated) gradient method (GD), i.e. $h(\mathbf{x}) = 0$ in (1), [16] showed that the gradient norm can attain a small value $\|\nabla f(\mathbf{x}_k)\|_2 \leq \epsilon$ in $\mathcal{O}\left(\frac{4L\|\mathbf{x}_0-\mathbf{x}^*\|_2}{\epsilon}\right)$ and $\mathcal{O}\left((\frac{L\|\mathbf{x}_0-\mathbf{x}^*\|_2}{\epsilon})^{2/3}\right)$ iterations for GD and its accelerated version (AGD) respectively. Similarly, other descent methods [17]–[19] such as stochastic gradient descent and sign gradient descent have successfully verified that the decrement of $\|\nabla f(\mathbf{x}_k)\|$ is not limited to ℓ_2 norm case.

For the ℓ_1 regularized problem, i.e. $h(\mathbf{x}) = \|\mathbf{x}\|_1$ in (1), the convergence of the gradient can be obtained from its dual problem, resulting in $\|\nabla f(\mathbf{x}^*)\|_\infty \leq \lambda$. In [20], it has been shown that this bound can be achieved after $\mathcal{O}\left(\kappa \cdot \log(\frac{2L\|\mathbf{x}_0-\mathbf{x}^*\|_2}{\varepsilon(\lambda,\nabla f(\mathbf{x}^*))})\right)$ iterations, where κ is the condition number of $f(\mathbf{x})$ and $\varepsilon(\cdot,\cdot)$ depends on λ (see (1)) and the entries of $|\nabla f(\mathbf{x}^*)|$. Additionally, it is known that $\|\nabla f(\mathbf{x}_k)\|_\infty$ is a sequence that tends to decrease, but not necessarily monotonously.

C. Robustifying FISTA via ℓ_∞ norm

The limit value of $\|\nabla f_k\|_\infty$ plays an important role in screening techniques [21] for quickly eliminating suboptimal features in the sparse solution. Recently, in the FISTA context, [8] has shown that $\|\nabla f_k\|_\infty$ can be employed as an early warning metric to identify future fluctuations in the objective. Furthermore, by forcing $\|\nabla f_k\|_\infty$ to be strictly decreasing ($\|\nabla f_k\|_\infty \geq \|\nabla f_{k+1}\|_\infty$) [8] heuristically showed that stability problems can be avoided.

Algorithm 1: Robustifying algorithm proposed in [8]

1 **if** $\|\nabla f_k\|_\infty > \tau_{k-1}$ **then**
2 Set $\nabla f_k(n) = \text{sign}(\nabla f_k(n)) \cdot \tau_{k-1}, \forall n \in \mathcal{I}$
3 $\tau_k = \tau_{k-1}, \quad c = \max(c_{\text{MIN}}, \rho \cdot c)$ (used in (5))
4 **else**
5 $\tau_k = \|\nabla f_k\|_\infty$

Such method can be broken down into an instability alert rule and a gradient correction, as shown in Algorithm 1, where $\mathcal{I} = \{n : |\nabla f_k(n)| > \tau_{k-1}\}$ is the set of "offending entries". Since the stability issues like divergence are caused by a too large SS, a multiplicative factor c as (5) can be used to reduce the SS, i.e. $\alpha_k = c \cdot \alpha_k$. In Algorithm 1, line 3, the factor $c = \max(c_{\text{MIN}}, \rho \cdot c)$ is gradually reduced ($\rho < 1$) until it reaches a predefined minimum value.

D. Over-relaxation and Under-relaxation

Relaxation technique is a versatile procedure that has been used in different scenarios [2], [22], [23] where it can enhance practical RoC at the cost of stability and viceversa by modifying the current estimation of a variable from the past estimation. The general relaxation model is given by

$$\mathbf{z}_{k+1} = \beta \mathbf{z}_{k+1} + (1-\beta)\mathbf{z}_k \tag{6}$$

where $\beta \in (0, 2)$ represents the relaxation factor. When the goal is to prioritize the convergence rate, β should be greater than 1, case that is known as over-relaxation and, for instance, it is used to accelerate the ADMM [2, Ch. 3] as well as other first order methods [23, Section 4]. On the other hand, if β is less than 1, the relaxation technique is called under-relaxation in which the modification of the current estimated variable can be interpreted as a correction that reinforces stability.

III. Proving convergence of the support gradient in FISTA method

On what follows, considering $h(\mathbf{x}) = \|\mathbf{x}\|_1$ in (1), we will assess the impact of replacing $\nabla f(\mathbf{y}_k)$ in (2) by

$$\nabla f_S(\mathbf{y}_k) = \text{supp}(\mathbf{y}_k) \odot \nabla f(\mathbf{y}_k), \tag{7}$$

i.e. gradient whose non-zero elements coincide with those of the current sparse solution support $\mathbf{y}_k$.

We start by expressing the gradient as

$$\nabla f(\mathbf{y}) = \nabla f_S(\mathbf{y}) + \nabla f_C(\mathbf{y}) \tag{8}$$

where $\nabla f_C(\mathbf{y})$ is the complementary vector such (7) holds. By construction, it is a direct exercise to note that $\langle \nabla f_S(\mathbf{y}), \nabla f_C(\mathbf{y})\rangle = \langle \mathbf{y}, \nabla f_C(\mathbf{y})\rangle = 0$. Furthermore, we also note that

$$\begin{aligned}\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{y})\rangle &= \langle \mathbf{x}-\mathbf{y}, \nabla f_S(\mathbf{y})\rangle + \langle \mathbf{x}-\mathbf{y}, \nabla f_C(\mathbf{y})\rangle \\ &= \langle \mathbf{x}-\mathbf{y}, \nabla f_S(\mathbf{y})\rangle + \langle \mathbf{x}, \nabla f_C(\mathbf{y})\rangle. \end{aligned} \tag{9}$$

Assuming a strict support shrinkage case, i.e. $\text{supp}(\mathbf{x}) \subseteq \text{supp}(\mathbf{y})$, where $\mathbf{x} = \mathbf{x}_{k+1}$ and $\mathbf{y} = \mathbf{x}_k$, then $\langle \mathbf{x}, \nabla f_C(\mathbf{y})\rangle$ would be zero since all non-zeros elements of $\nabla f_C(\mathbf{y})$ are canceled out by the zeros elements of $\mathbf{x}$. Thus, (9) is further simplified to

$$\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{y})\rangle = \langle \mathbf{x}-\mathbf{y}, \nabla f_S(\mathbf{y})\rangle \tag{10}$$

Finally by using (10) in the quadratic approximation of f, and after simple algebraic manipulation, we can obtain the following particular iterative sequence:

$$\begin{aligned}\mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}} \frac{1}{2\alpha}\|\mathbf{x} - (\mathbf{y} - \alpha\nabla f_S(\mathbf{y}))\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ &= \text{shrink}_{\alpha\lambda}(\mathbf{y} - \alpha\nabla f_S(\mathbf{y})) \end{aligned} \tag{11}$$

To guarantee that the new sequence (11) converges, the sufficient decrease lemma [24, Sect. 10.3] has to be proved. By the decrease equation [24, Sect. 5.1], where (10) is used, and the second prox theorem [24, Sect. 6.5], where $\mathbf{x} = \text{shrink}_{\alpha\lambda}(\mathbf{y} - \alpha\nabla f_S(\mathbf{y}))$, we have that

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f_S(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{1}{2\alpha_L}\|\mathbf{y}-\mathbf{x}\|_2^2 \tag{12}$$

$$\langle \mathbf{y} - \alpha \nabla f_S(\mathbf{y}) - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \leq \lambda\alpha \left(\|\mathbf{y}\|_1 - \|\mathbf{x}\|_1 \right) \tag{13}$$

where α_L is the inverse of the Lipschitz constant of f. Expressing (13) as $\langle \nabla f_S(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \lambda \left(\|\mathbf{y}\|_1 - \|\mathbf{x}\|_1 \right) - \frac{1}{\alpha}\|\mathbf{y} - \mathbf{x}\|_2^2$ and plugging such expression into (12), we obtain

$$F(\mathbf{y}) - F\big(\widetilde{T}_\alpha(\mathbf{y})\big) \geq \frac{\alpha(2\alpha_L - \alpha)}{2\alpha_L} \|\widetilde{G}_\alpha^F(\mathbf{y})\|_2^2 \tag{14}$$

where $F = f + \lambda\|\cdot\|_1$, $\widetilde{T}_\alpha^F(\mathbf{y}) = \text{shrink}_{\alpha\lambda}(\mathbf{y} - \alpha\nabla f_S(\mathbf{y}))$, $\widetilde{G}_\alpha^F(\mathbf{x}) = \frac{1}{\alpha}(\mathbf{y} - \widetilde{T}_\alpha^F(\mathbf{y}))$ and the required lemma is satisfied.

IV. Proposed method

Inspired by [8], summarized in Section II-C, we present an improved extension to such framework. Our proposed method also consists of an instability warning rule and a gradient correction stage, however, as described below, we modify both stages. A thorough experimental analysis[1] shows that in general $\|\nabla f_k\|_\infty$ decreases non-monotonically throughout iterations, even for conservative SS constant values. Thus, in order to account for such behavior, we proposed to include a trust bound in the instability warning rule, i.e. we check if

$$\|\nabla f_k\|_\infty > \mu \cdot \tau_{k-1}, \tag{15}$$

where $\mu > 1$ instead of using the original rule, which takes $\mu = 1$ (see line 1 in Algorithm 1).

Our experimental analysis[1] also showed that when an instability event occurs, there is a drastic alteration in a large number of ∇f_k's elements, which means an abrupt change in the direction or directional flow loss. In contrast to [8], which only changed the "offending entries" that triggered the activation of the warning rule, we propose to preserve its gradient direction as much as possible; this is described next.

Assuming that f is linear, then FISTA's extragradient rule (2) can be written as

$$\nabla f(\mathbf{y}_k) = \nabla f(\mathbf{x}_k + \gamma_k(\mathbf{x}_k - \mathbf{x}_{k-1})) \tag{16}$$

$$= (1 + \gamma_k)\nabla f(\mathbf{x}_k) - \gamma_k \nabla f(\mathbf{x}_{k-1}); \tag{17}$$

due to its structure, (17) can be considered as kind of over-relaxation applied to the gradient. As mentioned in Section II-D, it could promote stability problems. To deal with it, whenever an instability event occurs, we propose to apply an under-relaxation:

$$\nabla f(\mathbf{y}_k) = \beta \nabla f(\mathbf{x}_k) + (1 - \beta)\nabla f(\mathbf{x}_{k-1}) \tag{18}$$

where $\beta < 1$ is the under-relaxation parameter. Since (6) can be regrouped as in (16), it is worth mentioning that we are only altering the gradient $\nabla f(\mathbf{y}_k)$ but not the point at which it is evaluated.

As noted in [20], in general the current solution $\mathbf{x}_k$ tends to match the support of $\mathbf{x}^*$ after a given number of iterations. Thus if we consider (8), in this scenario, $\nabla f(\mathbf{y}_k)$ prioritizes the partial direction to the elements associated with the support of the sparse solution. Consequently, if an instability event occurs, we proposed to replace the current gradient by (7):

$$\nabla f(\mathbf{y}_k) = \nabla f_S(\mathbf{y}_k), \tag{19}$$

which can be interpreted as another type of under-relaxation. Finally, for this latter correction, our experimental analysis[1] also points out that it is necessary to preserve the support of the sparse solution in each iteration, thus the support of $\mathbf{y}_k$ is limited to that of $\mathbf{x}_k$, i.e. $\mathbf{y}_k = \text{supp}(\mathbf{x}_k) \odot \mathbf{y}_k$.

Algorithm 2: Proposed stability correction algorithm

1 **if** $\|\nabla f_k\|_\infty > \mu \cdot \tau_{k-1}$ **then**
2 Corrections: $\nabla f_k =$ (18) or (19)
3 $\tau_k = \tau_{k-1}, \quad c = \max(c_{\text{MIN}}, \rho \cdot c)$ (used in (5))
4 **else**
5 $\tau_k = \|\nabla f_k\|_\infty$

V. Computational Results

Our experiments were carried out on an Intel i7-7700HQ CPU (2.80 GHz, 6MB Cache, 8GB RAM). Analogous to [8], we evaluate our algorithms and others ones in the convolutional sparse coding (CSC) problem and the convolution dictionary learning (CDL) problem, (20) and (21) respectively:

$$\underset{\{\mathbf{x}_m\}}{\arg\min} \frac{1}{2} \sum_k \| \sum_m \mathbf{d}_m * \mathbf{x}_m - \mathbf{s}\|_2^2 + \lambda \sum_m \|\mathbf{x}_m\|_1, \tag{20}$$

$$\underset{\{\mathbf{x}_{k,m}\}\{\mathbf{d}_m\}}{\arg\min} \frac{1}{2} \sum_k \| \sum_m \mathbf{d}_m * \mathbf{x}_{k,m} - \mathbf{s}_k\|_2^2 + \lambda \sum_k \sum_m \|\mathbf{x}_{k,m}\|_1 \quad \text{s.t.} \quad \|\mathbf{d}_m\|_2 \leq 1 \;\; \forall m, \tag{21}$$

where $\mathbf{x}$ represents a feature map, $\mathbf{d}$ is a filter bank, and $\mathbf{s}$ is an observed image that in the CDL case is called training set.

The following robustifying stability algorithms are examined in this testing section

- **BaseCorr**: The robustifying algorithm as presented in [8].
- **ModBaseCorr**: Same as BaseCorr algorithm, but using a trust bound (15), with $\mu \neq 1$, as a warning rule.
- **PropCorr1**: The proposed algorithm (see Algorithm 2) using (18) as the gradient correction.
- **PropCorr2**: The proposed algorithm using (19) as the gradient correction.

The experiments are divided into two cases, the CSC and CDL problems solved via the FISTA method[2] (i) with an ill-chosen fixed SS and (ii) with the adaptive SS (5). Due to space constraints, we only assess the first case in the CSC context. For the CSC problem, we consider one different gray-scale observed image ("Bridge" and "Barbara") in each case. These images are corrupted with additive Gaussian noise $\sigma = 0.1$. For the CDL problem, we use 10 gray-scale training images of size 256×256 pixels, cropped and rescaled from a set of images obtained from the MIRFFLICKR-IM dataset [28]. The adjustable parameters used for both optimization problems and the robustifying stability algorithms are summarized in

[1]While not presented here due to space constrains, it can be fully reproduced via our companion software [25], [26]

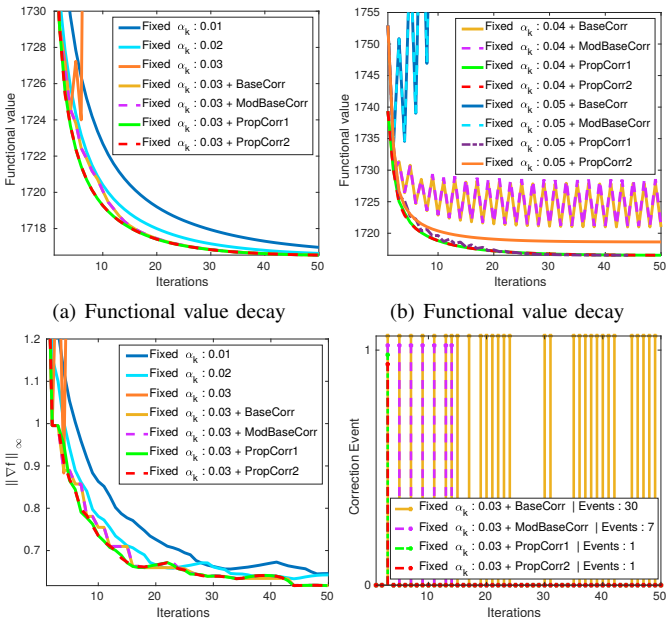
[2]The FISTA algorithms used to solve the CSC and CDL problems are extracted from [13] and [27].

Table I. Furthermore, we heuristically found that the optimal multiplicative factor μ of the trust bound[3] is 1.053 and 1.10 for PropCorr1 and PropCorr2 respectively, while the under-relaxation parameter β is equal to 0.95.

TABLE I: Summary of the parameters used in our experiments

Experiment	Problem	Image	λ	c_{MIN}	c	ρ
1. fixed SS	CSC	Bridge	0.6	1	1	1
2. adp. SS	CSC	Barbara	0.2	0.1	0.2 to 1	0.9
	CDL	10 training images	0.1	0.1	0.3 to 0.5	0.9

In the first group of experiments, we will focus on analyzing the contribution of the instability warning rule and the gradient correction stage by keeping constant the SS.



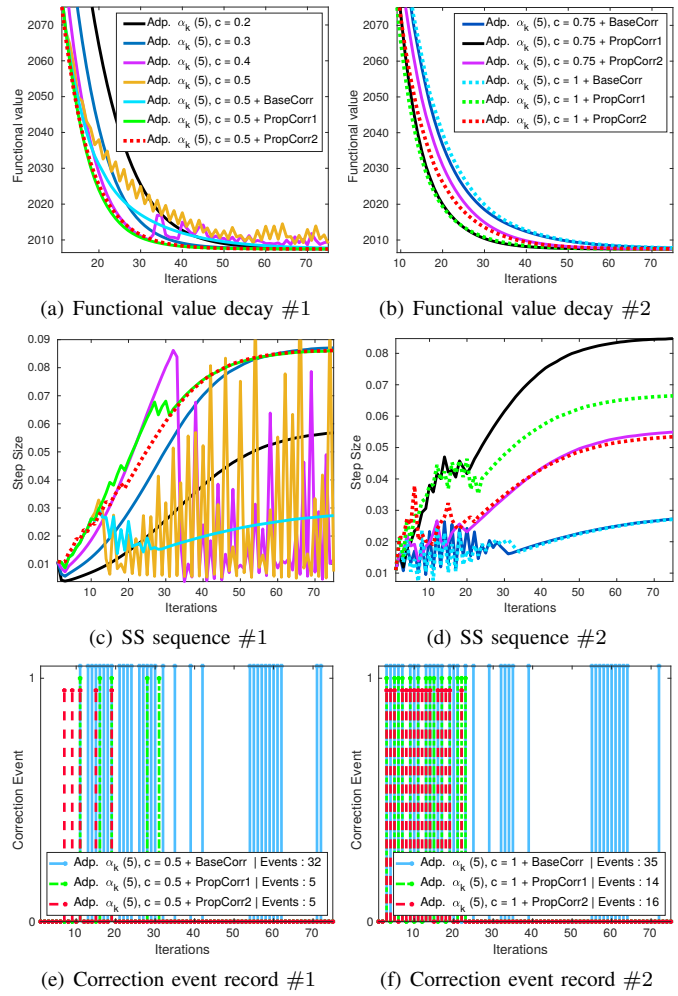
(a) Functional value decay (b) Functional value decay

(c) Evolution of $\|\nabla f_k\|_\infty$ (d) Correction event record

Fig. 1: CSC - A comparison of the FISTA with and without robustifying stability algorithms for several fixed SS values.

In Figures 1 (a)-(d), we report the progress of the function value and $\|\nabla f_k\|_\infty$, and the activation of stability correction events when solving the CSC problem via FISTA with fixed SS α_k from 0.01 to 0.05. As can be observed in the Figures 1(a) and 1(c), while the functional value of FISTA with fixed SS $\alpha_k = 0.01$ (blue line) converges monotonically, the evolution of $\|\nabla f_k\|_\infty$ is not monotonous, there are small increments between iteration 30 and iteration 50. Furthermore, it easy to note that a larger SS can provide a faster convergence, but also can generate stability issues (orange line). For the case of $\alpha_k = 0.03$, all considered algorithms can address the stability problems, and thereby achieve the same point of convergence. However, when comparing our warning rule for the same gradient correction (purple and yellow lines in the

[3]The multiplicative factor μ was related to the minimum number of ∇f_k's elements that can be greater in magnitude than τ_{k-1} without causing loss of directional flow.



(a) Functional value decay #1 (b) Functional value decay #2

(c) SS sequence #1 (d) SS sequence #2

(e) Correction event record #1 (f) Correction event record #2

Fig. 2: CSC - A comparison of the FISTA with and without robustifying stability algorithms for the adaptive SS (5).

Figure 1(d)), our rule helps to considerably reduce the number of correction events (whose computational cost isn't trivial), which are not really necessary since the base algorithm and our modification of it (purple and yellow lines) attains the same performance in terms of functional value decay with respect to number of iterations. On the other hand, when comparing the two variants of gradient correction with respect to the baseline (green/PropCorr1, red/PropCorr2 and yellow/BaseCorr lines in the Figures 1(a) and 1(d)), we can see that only a good gradient correction was needed to get the best convergence performance. Another characteristic to consider is that our proposed algorithms (PropCorr1 and PropCorr2) have a greater range of effectiveness that allows the usage of the larger step-sizes, see the Figures 1(a) and 1(b).

As mentioned before the adaptive SS (5) has a good performance, but it depends of an adequate selection of the multiplicative factor c. In Figure 2(a), we illustrate the experimental results for four values of $c = \{0.2,\ 0.3,\ 0.4,\ 0.5\}$, where the best value is $c = 0.3$, nevertheless, there is no guarantee that this value works in the same CSC problem with other settings. In CDL problem, each training image has a SS and therefore each SS should have an associated optimal value of

c to get the best performance, but in practice a single value is chosen. It can be seen in the Figures 2 and 3 that, for CSC and CDL problems, our algorithms (green and red lines) provide the best convergence behavior and allow to preserve a stable increasing SS sequences (behind both algorithms achieve good large c values for each image). As expected, the number of correction events of our algorithms is also much lower than the base algorithm (cyan line). Generally speaking, our two correction versions are competent, however, each one has an advantage over the other one. PropCorr1 is more generic, not limited to ℓ_1 regularized problems, and PropCorr2 is simpler in terms of computational complexity.

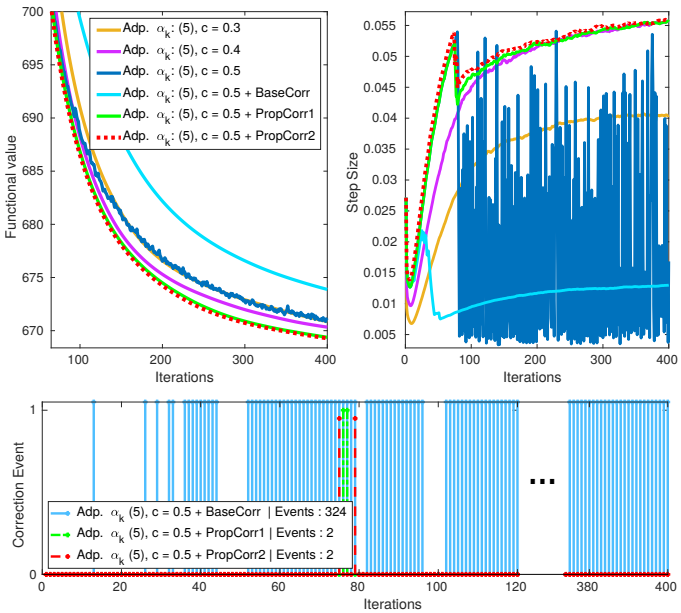


Fig. 3: CDL - A comparison of the FISTA with and without robustifying stability algorithms for the adaptive SS (5). The presented evolution of the SS and correction event are associated to one of the ten training image.

VI. Conclusions

FISTA is a gradient-based algorithm that can present stability issues (oscillation or divergence) when the SS is wrongly chosen or estimated. In this article, we have proposed an efficient algorithm to detect and correct potential stability issues in FISTA. Compared to its predecessor robustifying algorithm, for ℓ_1 regularized problems as CSC and CDL, our algorithm has shown to have a better instability warning rule that reduces possible false alerts. Furthermore, our adequate gradient corrections allowed to avoid future stability issues and obtained the best convergence performance.

References

[1] M. Schmidt, "Least squares optimization with ℓ_1-norm regularization," 2005.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[4] M. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM Journal on Numerical Analysis*, vol. 51, no. 2, pp. 927–957, 2013.

[5] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[6] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE TIP*, vol. 18, no. 11, pp. 2419–2434, Nov. 2009.

[7] J. Liang, J. Fadili, and G. Peyré, "Activity identification and local linear convergence of forward–backward-type methods," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 408–437, 2017.

[8] P. Rodriguez, "Robustifying fista via the ℓ_∞ norm of its smooth component's gradient," in *Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, 2020.

[9] F. Iutzeler and J. Malick, "On the proximal gradient algorithm with alternated inertia," *Journal of Optimization Theory and Applications*, vol. 176, no. 3, pp. 688–710, Mar 2018.

[10] P. Rodriguez, "Improving FISTA's speed of convergence via a novel inertial sequence," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[11] R. Gu and A. Dogandzić, "Projected nesterov's proximal-gradient algorithm for sparse signal recovery," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3510–3525, July 2017.

[12] M. Florea and S. Vorobyov, "A robust fista-like algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4521–4525.

[13] G. Silva and P. Rodriguez, "Fista: achieving a rate of convergence proportional to k^{-3} for small/medium values of k," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[14] Y. Yuan, "Step-sizes for the gradient method," *AMS IP Studies in Advanced Mathematics*, vol. 42, no. 2, p. 785, 2008.

[15] J. Barzilai and J. Borwein, "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, vol. 8, no. 1, pp. 141–148, 1988.

[16] Y. Nesterov, "How to make the gradients small," *Optima*, vol. 88, pp. 10–11, 2012.

[17] J. Nocedal, A. Sartenaer, and C. Zhu, "On the behavior of the gradient norm in the steepest descent method," *Computational Optimization and Applications*, vol. 22, pp. 5–35, 04 2002.

[18] Z. Allen-Zhu, "How to make the gradients small stochastically: Even faster convex and nonconvex sgd," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 1157–1167.

[19] L. Balles, F. Pedregosa, and N. Le Roux, "The geometry of sign gradient descent," *arXiv e-prints*, pp. arXiv–2002, 2020.

[20] J. Nutini, M. Schmidt, and W. Hare, ""active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern?" *Optimization Letters*, vol. 13, no. 4, pp. 645–655, 2019.

[21] L. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the lasso and sparse supervised learning problems," *CoRR*, vol. abs/1009.3515, 09 2010.

[22] J. Eckstein and D. Bertsekas, "On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 3, pp. 293–318, Jun. 1992.

[23] F. Iutzeler and J. Hendrickx, "A generic online acceleration scheme for optimization algorithms via relaxation and inertia," *Optimization Methods and Software*, vol. 34, no. 2, pp. 383–405, 2019.

[24] A. Beck, *First-Order Methods in Optimization*. SIAM, 2017.

[25] G. Silva, "Robustifying stability algorithms - matlab code," https://sites.google.com/a/pucp.edu.pe/gsilva/software.

[26] P. Rodriguez, "Simulations for first order methods - python code," https://gitlab.com/prodrig/f2o-master.

[27] G. Silva and P. Rodriguez, "Efficient convolutional dictionary learning using partial update fast iterative shrinkage-thresholding algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4674–4678.

[28] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 527–536.