Integral control of the proximal gradient method for unbiased sparse optimization

Integral control of the proximal gradient method for unbiased sparse optimization

V. Cerone, S. M. Fosson, A. Re, D. Regruto *

April 18, 2025

Abstract

Proximal gradient methods are popular in sparse optimization as they are straightforward to implement. Nevertheless, they achieve biased solutions, requiring many iterations to converge. This work addresses these issues through a suitable feedback control of the algorithm's hyperparameter. Specifically, by designing an integral control that does not substantially impact the computational complexity, we can reach an unbiased solution in a reasonable number of iterations. In the paper, we develop and analyze the convergence of the proposed approach for strongly-convex problems. Moreover, numerical simulations validate and extend the theoretical results to the non-strongly convex framework.

1 Introduction

Nowadays, parsimonious models, i.e., models that depend on a relatively small number of parameters, play a central role in machine learning, system identification and neural networks. While over-parametrization may have benefits in deep networks [1], parsimony is crucial for diverse purposes: it reduces the computational complexity for lightweight implementation, e.g., in mobile and cyber-physical applications; it provides interpretable representations of physical dynamical systems by selecting the most relevant variables; it prevents overfitting; it deals with compressed measurements and missing data. We refer the reader to, e.g., [2, 3, 4] for a comprehensive overview.

*The authors are with the Department of Control and Computer Engineering, Politecnico di Torino, Italy; e-mail: sophie.fosson@polito.it. Funded by the European Union - NextGenerationEU, Mission 4 Component 1.5 - ECS00000036 - CUP E13B22000020001.

In most cases, building parsimonious models from data consists in finding sparse solutions (i.e., solutions with many zeros) to minimization problems, which we refer to as *sparse optimization*. A valuable approach to promote sparsity and select the most important features is to add a regularization to the cost function to minimize. In the literature, considerable attention is devoted to ℓ_1 regularization, since the ℓ_1 norm is the best convex approximation of the number of non-zero components of a vector; see, e.g., [5].

Since sparsity-promoting regularization is usually non-differentiable, the proximal gradient method (PGM) is the natural alternative to gradient descent methods. PGM is an iterative algorithm that consists of a gradient step over the (differentiable) cost function and a proximal operator over the regularization; see, e.g., [6, 7, 8] for details. In case of ℓ_1 regularization, PGM is also known as iterative shrinkage-thresholding algorithm (ISTA, [9]), as the proximal map of the ℓ_1 norm shrinks and thresholds the current estimate.

A drawback of sparsity-promoting regularization is its inherently biased solution. The regularized problem calls for a tradeoff between minimization of the cost function and sparsity: usually, a satisfactory variable selection comes with an unavoidable inaccuracy in assessing the values of the selected variables.

The literature has devoted much attention to this issue for the Lasso estimator, i.e., for the ℓ_1-regularized least-squares minimization [5]. In particular, several works focus on non-convex regularization to correct the bias of Lasso; see, e.g., [10, 11, 12, 13]. Moreover, in [8], the authors notice that PGM applied to Lasso with non-convex regularization yields a faster convergence with respect to ISTA. This approach, denoted as AD-ISTA, has an adaptive shrinkage hyperparameter that speeds up the convergence while mitigating the bias effect. From a feedback control perspective, AD-ISTA is a discrete-time dynamical system whose shrinkage hyperparameter is a control input that evolves as a function of the current state value.

Starting from this general feedback perspective, this work proposes a novel approach to tune the hyperparameter of ISTA to keep the velocity of the adaptive approach while optimizing the bias control. More precisely, we investigate a control-theoretic approach and we design an integral control for ISTA by using the hyperparameter as a control input.

The contributions of the paper are twofold. Firstly, we develop the proposed approach, and we analyze its convergence for strongly convex cost functions. Secondly, we illustrate some numerical results to compare the proposed method and state-of-the-art gradient-based techniques in strongly

convex and non-strongly convex Lasso problems.

We organize the paper as follows. In Sec. 2, we state the problem. In Sec. 3, we develop the proposed approach and analyze its convergence in Sec. 4. Then, we validate and extend the theoretical results through numerical simulations in Sec. 5. Finally, we draw some conclusions.

2 Problem Statement

In this work, we consider optimization problems of the kind $\min_{x\in\mathbb{R}^n} f(x)$ where $f : \mathbb{R}^n \mapsto \mathbb{R}^+$ is convex, differentiable and admits a sparse minimizer that we aim to estimate.

Since the proposed study potentially addresses high-dimensional data problems, solving $\nabla f(x) = 0$ is not a viable way to estimate the sparse minimizer. Moreover, the problem is not well-posed when f has multiple minimizers. We resort to gradient-based methods for these motivations to achieve the desired solution.

To promote sparsity, we modify the problem into

$$\min_{x\in\mathbb{R}^n} f(x) + \sum_{i=1}^{n} \lambda_i |x_i| \tag{1}$$

where and $\sum_{i=1}^{n} \lambda_i |x_i|$ is a weighted ℓ_1 regularization with $\lambda_i \geq 0$ for each $i = 1, \ldots, n$.

If f admits a unique (sparse) minimizer, in principle, the sparsity-promoting regularization term is unnecessary and we can obtain the solution through gradient descent. However, this approach can be very slow; regularization improves the convergence rate at the price of a bias. On the other hand, if f admits multiple minimizers, the ℓ_1 regularization plays a crucial role to achieve the desired sparse estimate.

To solve (1), we can apply PGM, which iterates a gradient descent step over f with constant stepsize $\tau > 0$ and a proximal mapping with respect to the non-smooth regularization. In the case of ℓ_1 regularization, the proximal operator corresponds to the shrinkage-thresholding operator $\mathrm{S}_{\tau\lambda} : \mathbb{R}^n \mapsto \mathbb{R}^n$, which is defined componentwise as follows:

$$\begin{aligned} \mathrm{S}_{\tau\lambda_i}(z_i) &:= \operatorname*{argmin}_{x_i\in\mathbb{R}} \left[\tau\lambda_i |x_i| + \frac{1}{2}(x_i - z_i)^2\right], \quad z \in \mathbb{R} \\ &= \begin{cases} z_i - \operatorname{sign}(z_i)\tau\lambda_i & \text{if } |z_i| > \tau\lambda_i \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

In conclusion, PGM for (1), also known as ISTA, reads as follows: for $k = 0, 1, 2, \dots,$

$$x(k+1) = \mathrm{S}_{\tau\lambda}\left(x(k) - \tau\nabla f(x(k))\right). \tag{3}$$

When f is strongly convex, the map in (3) is contractive thanks to the non-expansiveness of $\mathrm{S}_{\tau\lambda}$, which implies convergence to the minimizer of (1), see, e.g., [14]. More generally, the convergence of PGM to a minimizer of (1) is studied, e.g., in [15, 6]. We remark that Lasso, as defined in [5], is an instance of problem (1) with $f(x) = \frac{1}{2}\|Ax - y\|_2^2$ where $A \in \mathbb{R}^{m,n}$, $y \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}^n$ has all equal components.

The main goal of this work is to tackle the following problem.

Problem 1. *Given ISTA as defined in* (3)*, we aim at designing a feedback control strategy, acting on λ as a control input, that minimizes the solution bias in* (1)*, that is, the distance from the minimum of f, while preserving the solution sparsity.*

2.1 Related literature

In the literature, two lines of research address the development of control strategies to improve the trajectory of ISTA for Lasso with acceleration purposes. The common idea is to design a time-varying λ that improves the dynamics of ISTA. As noticed in [16], for reasonably small $\lambda > 0$, ISTA for Lasso firstly minimizes f, then it adjusts the ℓ_1 norm; see [16, Fig.1]. This trajectory is not optimal because it causes a considerable ℓ_1 overshoot, whose correction is time consuming. To address this issue, the work [17] introduces D-ISTA, that is, an ISTA with geometrically decreasing $\lambda(k) > 0$. In other terms, in D-ISTA, λ can be interpreted as an open-loop control law. However, the design of a convenient control law is critical, as illustrated in [17, Theorem 3.2].

In contrast, in [8], the authors consider a feedback control approach for λ that originates from the use of a non-convex regularization instead of ℓ_1 norm. The corresponding PGM, called AD-ISTA, is usually faster than ISTA and other competitors in Lasso problems.

These two approaches enjoy an excellent interpretation in the framework of control methods, and the corresponding control laws derive from sparse optimization considerations. Moreover, their focus is on accelerating ISTA. In contrast, in Problem 1, we start from a control perspective and we focus on the bias regulation.

3 Proposed approach

As stated in Problem 1, the goal of this work is to remove the bias without affecting the solution sparsity. Since f is differentiable and convex by assumption, removing the bias corresponds to achieving $\nabla f = 0$. The key idea of the proposed approach is to regulate $y(k) = \nabla f(x(k))$ to zero by a suitable feedback control law on ISTA, with λ as a control input. A straightforward choice to regulate the output of a dynamical system to zero is to implement an integral control $\lambda(t) = k_i \int_0^t y(k)$, where $k_i \in \mathbb{R}$ is a design parameter. By recasting the integral law into a discrete-time setting, we obtain $\lambda(k+1) = \lambda(k) + k_i y(k)$, where k_i accounts also for the discretization step.

More precisely, the proposed ISTA with integral control, denoted by I-ISTA, is as follows:

$$\begin{cases} x(k+1) = \mathrm{S}_{\tau\lambda}\left(x(k) - \tau\nabla f(x(k))\right) \\ \lambda(k+1) = (1-\alpha)\lambda(k) + k_i \nabla f(x(k)) \end{cases} \tag{4}$$

where $\alpha \in (0,1)$ is a correction term useful for the convergence analysis presented in Sec. 4.

We notice that the computational complexity of I-ISTA is similar to that of ISTA because the two algorithms differ only for the update of λ in (4). Regarding the storage requirements, I-ISTA requires to save $2n$ variables instead of n at each iteration.

4 Convergence analysis

In this section, we characterize the equilibrium point of (4) in a strongly convex framework and we prove the convergence of I-ISTA.

We consider the following assumptions.

Assumption 1. *f is differentiable and μ-strongly convex, i.e., there exists $\mu > 0$ such that $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.*

This implies that, for any $x, z \in \mathbb{R}^n$,

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z) + \frac{\mu}{2}\|x - z\|_2^2. \tag{5}$$

Assumption 2. *f is β-smooth, that is, there exists $\beta > 0$ such that $\frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex.*

β-smoothness corresponds to the β-Lipschitz continuity of ∇f. Thus, for any $x, z \in \mathbb{R}^n$

$$f(x) \leq f(z) + \nabla f(z)^\top (x - z) + \frac{\beta}{2}\|x - z\|_2^2. \tag{6}$$

We refer the reader to, e.g., [18] for details.

The following result states that the equilibrium point of system (4) provides an unbiased solution.

Lemma 1. *Let Assumption 1 holds. If $\alpha > |k_i|$, the equilibrium point $(x^\star, \lambda^\star)$ of (4) satisfies $\nabla f(x^\star) = 0$ and $\lambda^\star = 0$. In particular, $x^\star$ is the unique, hence sparse minimizer of f.*

Proof. We compute the equilibrium points of (4).

$$\begin{cases} x^\star = \mathrm{S}_{\tau\lambda^\star}\left(x^\star - \tau\nabla f(x^\star)\right) \\ 0 = -\alpha\lambda^\star + k_i \nabla f(x^\star) \end{cases} \tag{7}$$

From the second equation, we have

$$\nabla f(x^\star) = \frac{\alpha}{k_i}\lambda^\star. \tag{8}$$

By replacing (8) in the first equation of (7),

$$x^\star = \mathrm{S}_{\tau\lambda^\star}\left(x^\star - \tau\frac{\alpha}{k_i}\lambda^\star\right). \tag{9}$$

Now, for each $j \in \{1, \ldots, n\}$ such that $x_j^\star \neq 0$,

$$x_j^\star = x_j^\star - \tau\frac{\alpha}{k_i}\lambda_j^\star - \operatorname{sign}\left(x_j^\star - \tau\frac{\alpha}{k_i}\lambda_j^\star\right)\tau\lambda_j^\star. \tag{10}$$

Therefore, $\lambda_j^\star = 0$ if $\frac{\alpha}{|k_i|} \neq 1$. On the other hand, for each $j \in \{1, \ldots, n\}$ such that $x_j^\star = 0$,

$$\left|\tau\frac{\alpha}{k_i}\lambda_j^\star\right| \leq \tau\lambda_j^\star. \tag{11}$$

If $\frac{\alpha}{|k_i|} > 1$, then (11) holds if $\lambda_j^\star = 0$.

In conclusion, $\lambda^\star = 0$, which implies $\nabla(f(x^\star)) = 0$ from (8). Moreover, $x^\star$ is the unique minimizer of f since $\nabla(f(x^\star)) = 0$ and f is μ-strongly convex. □

Now, let us analyze the convergence of I-ISTA to the equilibrium point described in Lemma 1.

Proposition 1. *Let us set* $\tau < \frac{2}{\beta}$ *and let*

$$\xi^2 = \max\{\sigma^2 + k_i^2\beta^2, \tau^2 + (1-\alpha)^2\} < \frac{1}{2}. \tag{12}$$

If Assumptions 1 and 2 hold, each step of I-ISTA is a contractive map, and I-ISTA converges to $(x^\star, \lambda^\star)$.

Proof. As illustrated in [19], since S is non-expansive, given assumptions 1-2 for any $x, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^n_+$

$$\|\mathrm{S}_{\tau\lambda}(x - \tau\nabla f(x)) - \mathrm{S}_{\tau\lambda}(z - \tau\nabla f(z))\|_2^2 \leq \sigma^2\|x - z\|_2^2 \tag{13}$$

where $\sigma^2 = \max\{(1-\tau\mu)^2, (1-\tau\beta)^2\}$. Since $\mu \leq \beta$, if $\tau < \frac{2}{\beta}$ then $\sigma^2 \in (0,1)$, that is, $h(x) = \mathrm{S}_{\tau\lambda}(x - \tau\nabla f(x))$ is contractive; see [19, Proof of Lemma 1].

On the other hand, for any $\lambda, \gamma \in \mathbb{R}^n_+$ and $x \in \mathbb{R}^n$

$$\|\mathrm{S}_{\tau\lambda}(x) - \mathrm{S}_{\tau\gamma}(x)\|_2^2 \leq \tau^2\|\lambda - \gamma\|_2^2. \tag{14}$$

In fact, through componentwise analysis,

1. if $|x_i| > \tau\lambda_i > \tau\gamma_i$, then $\mathrm{S}_{\tau\lambda_i}(x_i) - \mathrm{S}_{\tau\gamma_i}(x_i) = \mathrm{sign}(x_i)\tau(\gamma_i - \lambda_i)$;
2. if $\tau\lambda_i > |x_i| > \tau\gamma_i$, then $|\mathrm{S}_{\tau\lambda_i}(x_i) - \mathrm{S}_{\tau\gamma_i}(x_i)| = |x_i - \mathrm{sign}(x_i)\tau\gamma_i| \leq \tau|\lambda_i - \gamma_i|$.

By using (13) and (14) and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, for any $(x, \lambda) \in \mathbb{R}^{2n}$ and $(z, \gamma) \in \mathbb{R}^{2n}$, we conclude

$$\begin{aligned} &\|\mathrm{S}_{\tau\lambda}(x - \tau\nabla f(x)) - \mathrm{S}_{\tau\gamma}(z - \tau\nabla f(z))\|_2^2 \leq \\ &2\,\|\mathrm{S}_{\tau\lambda}(x - \tau\nabla f(x)) - \mathrm{S}_{\tau\lambda}(z - \tau\nabla f(z))\|_2^2 + \\ &2\,\|\mathrm{S}_{\tau\lambda}(z - \tau\nabla f(z)) - \mathrm{S}_{\tau\gamma}(z - \tau\nabla f(z))\|_2^2 \leq \\ &2\sigma^2\|x - z\|_2^2 + 2\tau^2\|\lambda - \gamma\|_2^2. \end{aligned} \tag{15}$$

As to the first equation of (4), the bound (15) implies that

$$\|x(k+1) - x^\star\|_2^2 \leq 2\sigma^2\|x(k) - x^\star\|_2^2 + 2\tau^2\|\lambda(k) - \lambda^\star\|_2^2. \tag{16}$$

Furthermore, regarding the second equation of (4), we notice that for any $(x, \lambda) \in \mathbb{R}^{2n}$ and $(z, \gamma) \in \mathbb{R}^{2n}$

$$\begin{aligned} &\|(1-\alpha)\lambda + k_i \nabla f(x) - (1-\alpha)\gamma - k_i \nabla f(z)\|_2^2 \\ &\leq 2(1-\alpha)^2 \|\lambda - \gamma\|_2^2 + 2k_i^2 \|\nabla f(x) - \nabla f(z)\|_2^2 \\ &\leq 2(1-\alpha)^2 \|\lambda - \gamma\|_2^2 + 2k_i^2 \beta^2 \|x - z\|_2^2 \end{aligned} \tag{17}$$

where in the last step we exploit the β-smoothness of f. Thus,

$$\begin{aligned} &\|\lambda(k+1) - \lambda^\star\|_2^2 \\ &\leq 2(1-\alpha)^2 \|\lambda(k) - \lambda^\star\|_2^2 + 2k_i^2 \beta^2 \|x(k) - x^\star\|_2^2. \end{aligned} \tag{18}$$

By summing (16) and (18), we obtain

$$\begin{aligned} &\|x(k+1) - x^\star\|_2^2 + \|\lambda(k+1) - \lambda^\star\|_2^2 \\ &\leq 2\xi^2 \left(\|x(k) - x^\star\|_2^2 + \|\lambda(k) - \lambda^\star\|_2^2 \right) \end{aligned} \tag{19}$$

where $\xi^2 = \max\{\sigma^2 + k_i^2\beta^2, \tau^2 + (1-\alpha)^2\} < \frac{1}{2}$. This proves that the mapping between $(x(k), \lambda(k))$ and $(x(k+1), \lambda(k+1))$ is contractive with coefficient $2\xi^2$, thus I-ISTA converges to $(x^\star, \lambda^\star)$ thanks to the Banach fixed-point theorem [20]. □

Remark 1. *The sufficient conditions of Proposition 1 are quite restrictive: to guarantee the contractivity of I-ISTA, we exploit the μ-strong convexity of f and the bound* (12)*, which limits the values of α. However, the obtained bounds are not tight, and numerical results prove that the conditions can be relaxed; see Sec. 5.*

5 Numerical results

In this section, we present some numerical results that support the theoretical convergence results and extend them to non-strongly convex problems. Moreover, they provide more insight into the trajectory and convergence speed of I-ISTA with respect to state-of-the-art gradient-based approaches.

Specifically, we compare I-ISTA to ISTA and its fast version, FISTA, introduced by [21], and to AD-ISTA and its fast version, AD-FISTA; see [8]. As shown in [8], in some Lasso problems, AD-FISTA is the fastest algorithm among state-of-the-art iterative sparse optimization algorithms. Moreover, we show the behavior of the gradient descent (without sparsity promoting terms) as a benchmark.

Our experiments focus on recovering sparse vectors from linear measurements via Lasso. We consider $\widetilde{x} \in \mathbb{R}^n$ with $n = 200$ and sparsity $\|\widetilde{x}\|_0 = 10 \ll n$. We randomly generate the non-zero components of $\widetilde{x}$ through a uniform distribution, with magnitude in $(1, 2)$. We aim to recover $\widetilde{x}$ from $y = A\widetilde{x}$, where $A \in \mathbb{R}^{m,n}$ has components independently generated with Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$. The cost function is $f(x) = \frac{1}{2}\|Ax-y\|_2^2$. We envisage either strongly convex ($m = 210 > n$) and non-strongly convex ($m = 150 < n$) cases.

To implement ISTA and FISTA, we consider Lasso with $\lambda_i = 10^{-3}$ for each $i = 1, \ldots, n$. For AD-ISTA and AD-FISTA, we consider a Log-Lasso with initial $\lambda_i = 3 \times 10^{-3}$ for each $i = 1, \ldots, n$ and $\epsilon = 10^{-2}$. The gradient stepsize is $\tau = \|A\|_2^{-2}$ for all the algorithms. For I-ISTA, we set $k_i = 10^{-3}$, while $\alpha = 0.05$ for $m = 210$ and $\alpha = 0.02$ for $m = 150$. For all the algorithms, the stop criterion is $\|x(k+1)-x(k)\|_2 < 10^{-10}$, with a maximum of iterations set to 5×10^4.

5.1 Strongly convex case: $m = 210 > n$

In Fig. 1, we show the trajectories of all the considered algorithms in the plane $\|Ax(k) - y\|_2$ vs $\|x(k)\|_1$. In contrast to existing ISTA-based approaches, the proposed I-ISTA converges to the true vector $\widetilde{x}$ without bias. The required number of iterations is comparable to ISTA and FISTA. We remark that the number of iterations is a valuable performance metric because the computational complexity of each iteration step is comparable for all the considered algorithms. The gradient descent (GRAD) is also unbiased, although the convergence is very slow.

A remarkable point is the linear trajectory of I-ISTA in the plane $\|Ax(k)-y\|_2$ vs $\|x(k)\|_1$, which highlights an optimal balance between decreasing the residual and increasing the ℓ_1 norm.

To substantiate these findings, in Fig. 1, we illustrate the time evolution of the relative error $\|x(k) - \widetilde{x}\|_2/\|\widetilde{x}\|_2$ and of the residual $\|Ax(k) - y\|_2$, averaged over 100 different runs. Finally, we show the evolution of the estimated support, i.e., the set of non-zero components, in Fig. 3. We notice that I-ISTA identifies the correct support; in fact, the support error defined as $\sum_{i=1}^n |1(x_i(k) - 1(\widetilde{x}_i)|$, where 1 denotes the indicator function $1(z) = \|z\|_0$ for $z \in \mathbb{R}$, goes to zero in all the runs. Moreover, I-ISTA identifies the correct support after a number of iterations comparable to AD-ISTA and AD-FISTA. In Fig. 3 (right), we highlight a peculiarity of I-ISTA: in contrast to the competitors, it builds the support from below, without the usual "false positives" phase that characterizes the ISTA-based

methods. This feature can be of interest for all those applications where the transient overestimated support can cause serious false alarms; this is the case, for example, of secure state estimation problems in cyber-physical systems, where the support denotes a subset of sensors under malicious attacks; see, e.g., [22] for details.

5.2 Non-strongly convex case: $m = 150 < n$

In this section, we duplicate the numerical simulations with $m = 150$. The primary outcome is that I-ISTA converges also for non-strongly convex problems, extending the landscape concerning the proposed convergence analysis.

In this case, f has infinitely many minimizers, and the gradient descent is ineffective because it does not converge to the sparse solution. In contrast, I-ISTA converges precisely to the desired solution, as we can see in Fig. 4 and 5. As in the strongly convex case, the convergence time of I-ISTA is comparable to ISTA and FISTA, while the support stabilization time, depicted in Fig. 6, is comparable to AD-ISTA and AD-FISTA.

	$m = 210$		$m = 150$	
Algorithm	Conv.	Supp. stab.	Conv.	Supp. stab.
GRAD	47183.57	8002.83	2687.42	–
ISTA	486.36	382.36	1761.47	1617.16
FISTA	322.40	255.76	1172.71	1079.11
AD-ISTA	123.80	23.70	172.80	45.07
AD-FISTA	80.01	16.70	113.17	31.09
I-ISTA	426.33	8.23	1107.80	25.40

Table 1: Mean number of iterations to converge and to stabilize the support, over 100 random runs.

In Table 1, we summarize the data about convergence and support stabilization times.

6 Conclusions

In this work, we propose and analyze I-ISTA, a proximal gradient method for ℓ_1-regularized sparse optimization problems with an integral control that removes the bias. We analyze the convergence of the algorithm in the framework of strongly convex problems, while numerical results extend its validity to non-strongly convex problems. The "linear" trajectory of I-ISTA yields a fast stabilization of the support estimate without support overshoot. Significant extensions under investigation are the convergence analysis in non-

strongly convex frameworks and the robustness to noise. Furthermore, we are studying how to expand and refine the approach by considering controllers that are more sophisticated than the integral one.

References

[1] G. Pillonetto, A. Aravkin, D. Gedon, L. Ljung, A. H. Ribeiro, and T. B. Schön, "Deep networks for system identification: A survey," *Autom.*, vol. 171, p. 111907, 2025.

[2] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[3] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, 2nd ed. CRC press, 2015.

[4] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

[5] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Stat. Soc. Series B*, vol. 58, pp. 267–288, 1996.

[6] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*. Springer New York, 2011, pp. 185–212.

[7] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[8] V. Cerone, S. M. Fosson, and D. Regruto, "Fast sparse optimization via adaptive shrinkage," *IFAC-PapersOnLine - IFAC World Congress*, vol. 56, no. 2, pp. 10 390–10 395, 2023.

[9] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413 – 1457, 2004.

[10] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[11] J. Woodworth and R. Chartrand, "Compressed sensing recovery via nonconvex shrinkage penalties," *Inverse Problems*, vol. 32, no. 7, pp. 75 004–75 028, 2016.

[12] A. Rakotomamonjy, G. Gasso, and J. Salmon, "Screening rules for Lasso with non-convex sparse regularizers," in *Proc. Int. Conf. Machine Learn. (ICML)*, 2019, pp. 5341–5350.

[13] S. Fosson, V. Cerone, D. Regruto, and T. Abdalla, "A concave approach to errors-in-variables sparse linear system identification," *IFAC-PapersOnLine - SYSID*, vol. 54, no. 7, pp. 298–303, 2021.

[14] M. Fornasier, "Numerical methods for sparse recovery," in *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Radon Series Comp. Appl. Math., de Gruyter, 2010, pp. 93–200.

[15] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality," *Math. Operations Res.*, vol. 35, no. 2, pp. 438–457, 2010.

[16] I. Daubechies, M. Fornasier, and I. Loris, "Accelerated projected gradient method for linear inverse problems with sparsity constraints," *J. Fourier Anal. Appl.*, vol. 14, pp. 764–792, 2008.

[17] S. Dahlke, M. Fornasier, and T. Raasch, "Multilevel preconditioning and adaptive sparse solution of inverse problems," *Math. Comput.*, vol. 81, no. 277, pp. 419–446, 2012.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[19] S. Hassan-Moghaddam and M. R. Jovanović, "Proximal gradient flow and Douglas–Rachford splitting dynamics: Global exponential stability via integral quadratic constraints," *Autom.*, vol. 123, p. 109311, 2021.

[20] S. Banach, "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales," *Fund. Math.*, vol. 3, pp. 133–181, 1922.

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

[22] V. Cerone, S. Fosson, D. Regruto, and F. Ripa, "Lasso-based state estimation for cyber-physical systems under sensor attacks," *IFAC-PapersOnLine - SYSID*, vol. 58, no. 15, pp. 163–168, 2024.

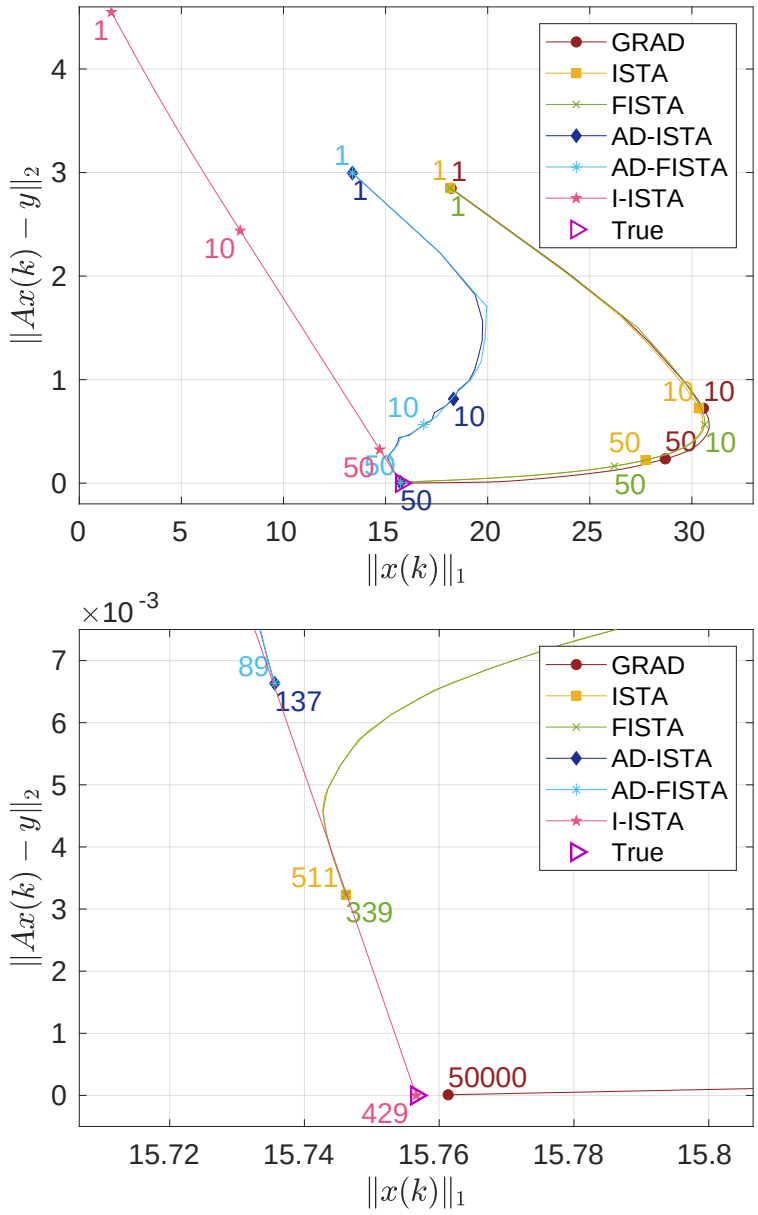

Figure 1: Example 1: $m = 210$. Residual $\|Ax(k) - y\|_2$ with respect to $\|x(k)\|_1$ in a single run. The curves are parametrized with time. "True" refers to the value of $\widetilde{x}$. On the left, we show the overall trajectory; we label iterations 1, 10, 50. On the right, we magnify the figure around $\widetilde{x}$ and report the convergence step. The gradient descent (GRAD) reaches $\widetilde{x}$, but with a number of iterations larger than the set maximum 5×10^4.

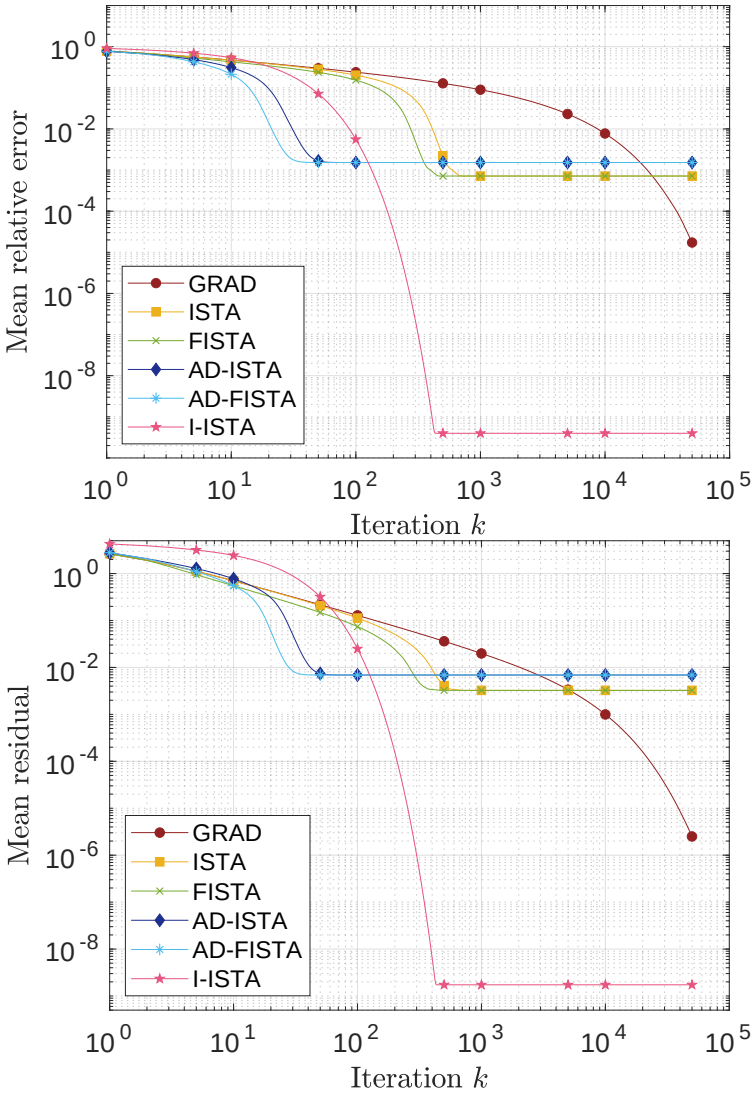

Figure 2: Example 1: $m = 210$. Evolution of the relative error $\|x(k) - \widetilde{x}\|_2/\|\widetilde{x}\|_2$ (left) and of the residual $\|Ax(k) - y\|_2$ (right), averaged over 100 runs.

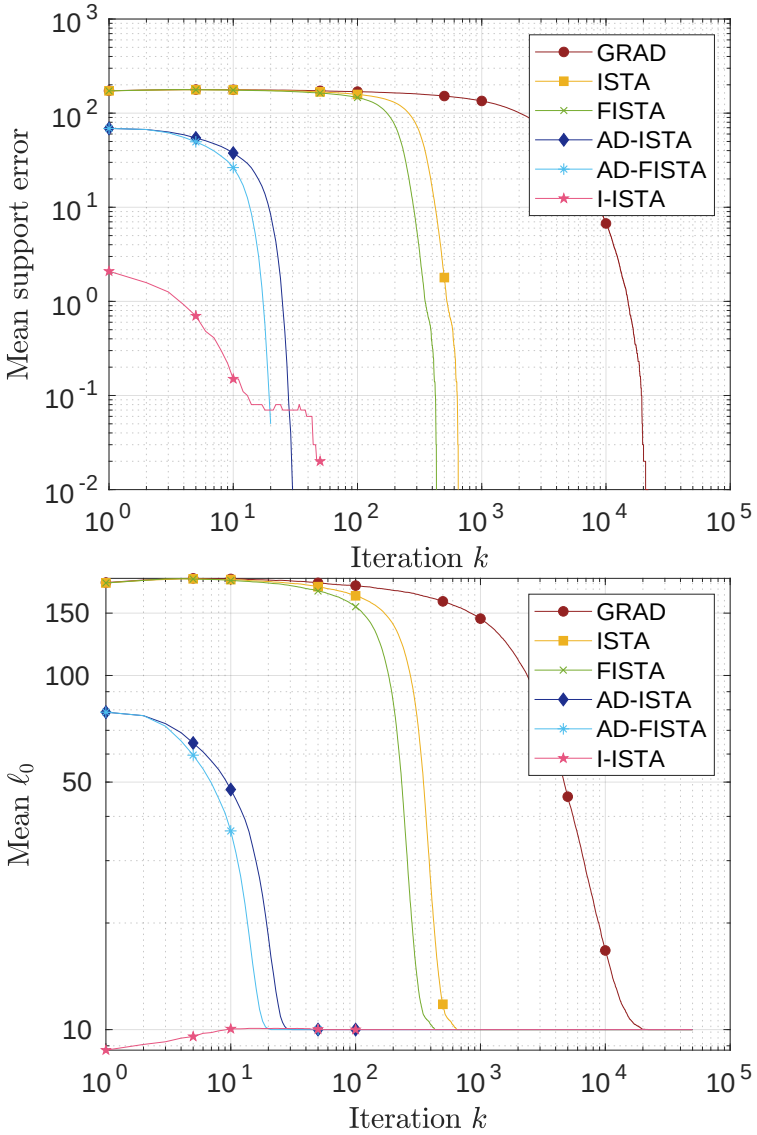

Figure 3: Example 1: $m = 210$. Evolution of the support error $\sum_{i=1}^{n} |1(x_i(k) - 1(\widetilde{x}_i)|$ (left) and of the sparsity level $\|x(k)\|_0$ (right), averaged over 100 runs. The graphs on the support error are interrupted when the error is null.

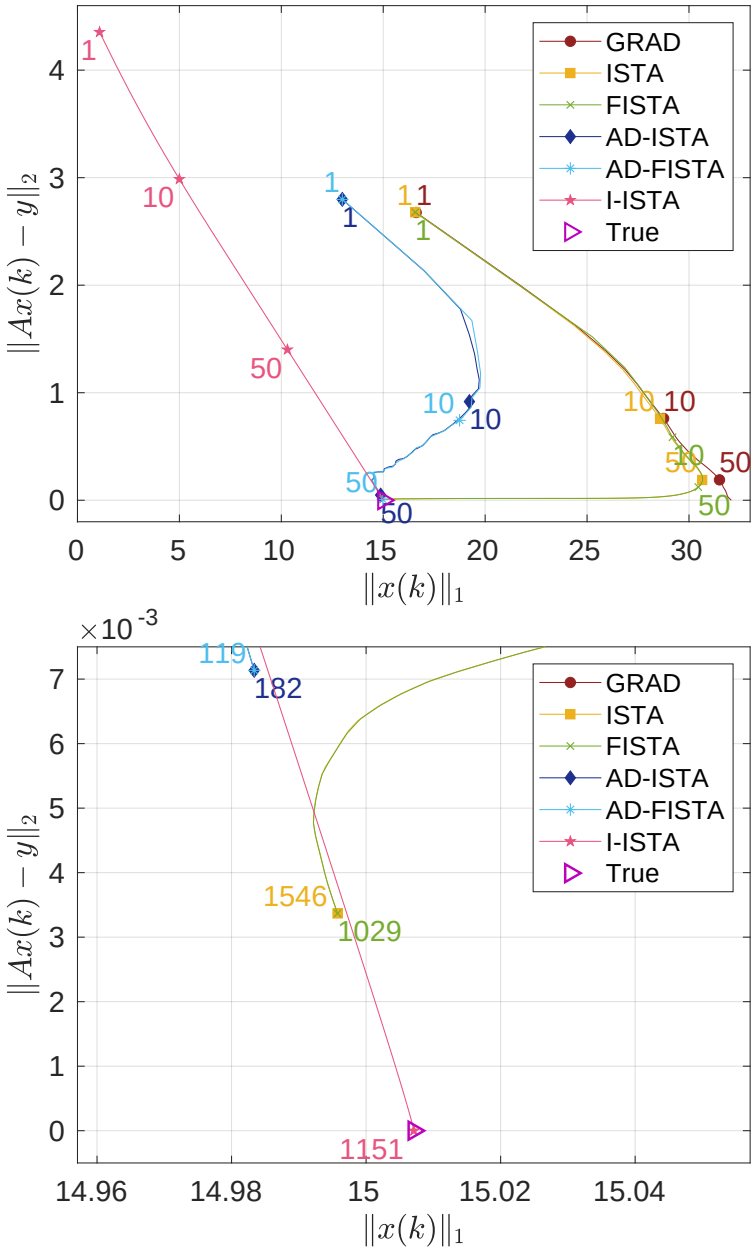

Figure 4: Example 2: $m = 150$. Residual $\|Ax(k) - y\|_2$ with respect to $\|x(k)\|_1$ in a single run. The curves are parametrized with time. "True" refers to the value of $\widetilde{x}$. On the left, we show the overall trajectory; we label iterations 1, 10, 50. On the right, we magnify the figure around $\widetilde{x}$ and report the convergence step.

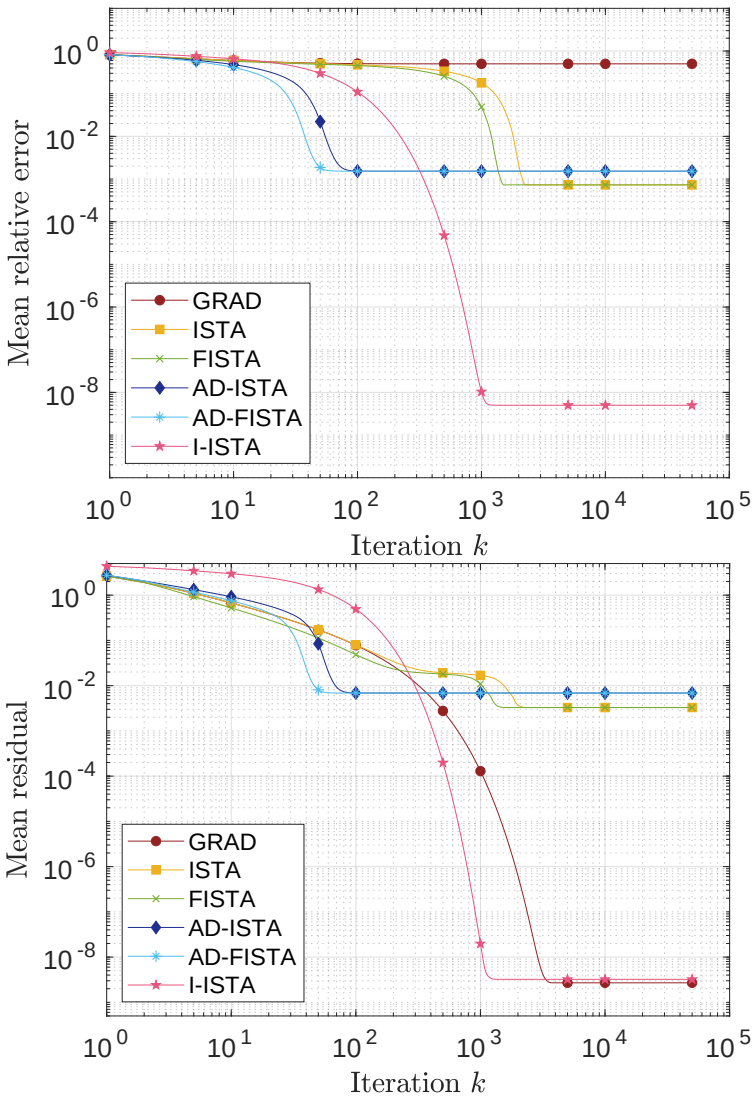

Figure 5: Example 2: $m = 150$. Evolution of the relative error $\|x(k) - \widetilde{x}\|_2/\|\widetilde{x}\|_2$ (left) and the residual $\|Ax(k) - y\|_2$ (right), averaged over 100 runs.

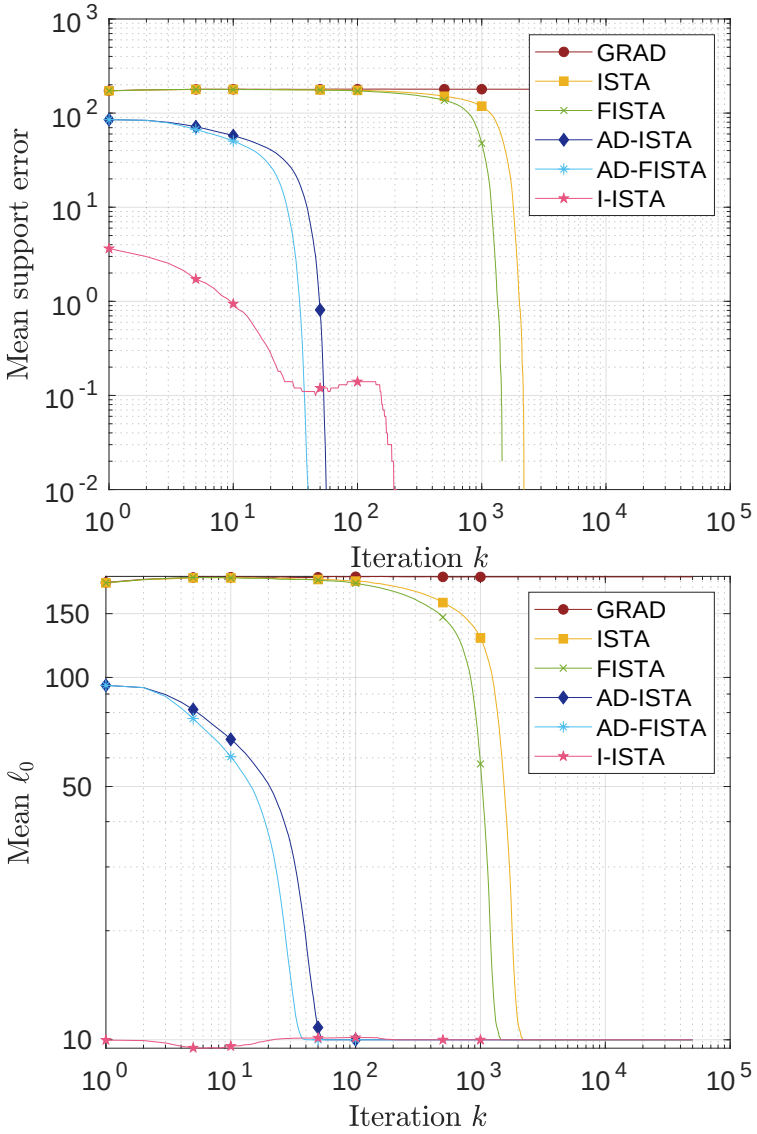

Figure 6: Example 2: $m = 150$. Evolution of the support error $\sum_{i=1}^{n} |1(x_i(k) - 1(\widetilde{x}_i)|$ (left) and of the sparsity level $\|x(k)\|_0$ (right), averaged over 100 runs. The graphs on the support error are interrupted when the error is null.