ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Revisiting FISTA for Lasso: Acceleration Strategies Over The Regularization Path

Alejandro Catalina, Carlos M. Alaíz and José R. Dorronsoro *

Departamento de Ingeniería Informática e Instituo de Ingeniería del Conocimiento
Universidad Autónoma de Madrid, Madrid, Spain

Abstract. In this work we revisit FISTA algorithm for Lasso showing that recent acceleration techniques may greatly improve its basic version, resulting in a much more competitive procedure. We study the contribution of the different improvement strategies, showing experimentally that the final version becomes much faster than the standard one.

1 Introduction

The growing popularity of Big Data and the corresponding increasingly larger problems in Machine Learning have led to a significant focus on sparse linear models such as Lasso [1]. For centered data Lasso can be written as

$$\min_{\beta\in\mathbb{R}^d} f(\beta) = \frac{1}{2N}\left\|X\beta - y\right\|_2^2 + \lambda\left\|\beta\right\|_1, \tag{1}$$

with $X \in \mathbb{R}^{N\times d}$ the data matrix, $y \in \mathbb{R}^N$ the target vector, $\beta \in \mathbb{R}^d$ the Lasso coefficients and the subscripts 1 and 2 denote the ℓ_1 and ℓ_2 norms respectively. There are two major approaches to solve Lasso. The first one is cyclic coordinate descent as implemented in the GLMNet algorithm [2], which is currently considered as the state-of-the-art. While GLMNet deals with a coefficient β_i at a time, its alternative, FISTA algorithm [3], goes in some sense to the other extreme, updating all the d components of β at each step by combining proximal gradient descent with Nesterov's accelerations.

Although it can be seen as a more general optimization algorithm, FISTA is not currently considered as competitive for Lasso. In contrast with the cheap and exact single coordinate iterations of GLMNet, two possible drawbacks of the full coordinate FISTA iterations stand out. The first one is their non-monotonic nature, due to an overshooting in Nesterov's momentum term [4] which results in a characteristic rippling behaviour for the objective function f. The other is the difficulty of computing sharp enough values of f's Lipschitz constant, as the standard backtracking strategy often results in too conservative, large estimates and, hence, in shorter, less effective gradient steps.

*Work with partial support from Spain's grants TIN2013-42351-P, TIN2016-76406-P, TIN2015-70308-REDT and S2013/ICE-2845 CASI-CAM-CM. Also supported by project FACIL–Ayudas Fundación BBVA a Equipos de Investigación Científica 2016, and the UAM–ADIC Chair for Data Science and Machine Learning. We thank Red Eléctrica de España for making available wind energy data and gratefully acknowledge the use of the facilities of Centro de Computación Científica (CCC) at UAM.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

A simple solution for the rippling behaviour is to restart Nesterov's momentum coefficient sequence when non-monotonicity is observed [4]. More generally, in a recent contribution Ito *et al.* [5] have put together a set of accelerations for FISTA in classification problems that address both drawbacks and suggest it may be worthwhile to revisit it as a more competitive algorithm for Lasso and other composite problems. This is the goal of this work, in which we adapt Ito's techniques to regression; our main contributions are:

- A grouping of the strategies in [5] as two main increasing algorithmic improvements over FISTA plus backtracking, dealing first with Nesterov's momentum and, second, improving the Lipschitz constant estimation in backtracking.
- The detailed study of the contribution of each acceleration strategy and their effects over an entire regularization path instead of the single λ executions in [5].

While we will only measure the number of iterations needed, the final improved version of FISTA is clearly more efficient, suggesting it to be worthwhile a further study of ways of increasing its competitiveness. The remaining of the paper is organized as follows. In Section 2 we briefly review FISTA and describe the acceleration strategies we consider. We present our experimental comparison in Section 3 and in Section 4 we offer other insights and pointers to further work.

2 FISTA and its Acceleration

2.1 Basic FISTA

FISTA (which stands for Fast Iterative Soft-Thresholding Algorithm) is an iterative algorithm based on the application of proximal operators to solve composite problems. For least squares it combines the basic iterations of ISTA [3]

$$\beta_k = S_{\frac{1}{L}}(w_k - \frac{1}{L}((X^T X + \mu I)w_k - X^T y)), \tag{2}$$

where $S_\gamma(z) = \text{sign}(z)(|z| - \gamma)_+$ and L is an estimate for the Lipschitz constant of the problem, with a Nesterov step:

$$w_{k+1} = \beta_k + \frac{t_k - 1}{t_{k+1}}(\beta_k - \beta_{k-1}), \;\; t_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right),$$

which adds a momentum term defined by the increasing t_k sequence. FISTA constitutes a generic algorithm with guaranteed convergence to the global minimum and that can be used in many problems, Lasso among the best known. For more details refer to the original paper [3].

Nonetheless, while FISTA's generic implementation may be an advantage over other problem-specific methods, it is not currently regarded as a state-of-the-art method to solve Lasso. One reason is the non-monotonicity caused by

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Nesterov's step. In fact, since the momentum terms grows per iteration, we may reach a point where we exceed its optimal value, getting a rippling behaviour that severely impacts performance. Another reason is its gradient step size, which may be too small sometimes. There have been several recent contributions that try to avoid these effects and suggest to reconsider FISTA as an efficient option for Lasso. In particular, Ito *et al.* [5], building on the work in [3, 4, 6], propose a new algorithm named Fast Accelerated Proximal Gradient which puts together several acceleration strategies to address the previous FISTA drawbacks. We briefly describe these strategies next.

2.2 Bactracking

As Beck and Teboulle already explained in their paper [3], in many cases we do not know the exact value of the Lipschitz constant of a problem and thus we are forced to estimate a suitable one L_k at each step. The backtracking strategy does precisely this, improving on the global Lipschitz constant L, which generally is too conservative, and yielding better estimates L_k that allows us to achieve a faster convergence. Backtracking guarantees the sequence L_k to be non-decreasing, something needed to fulfil the conditions required for the improved convergence analysis in [3].

2.3 Restarting and Maintaining Top Speed

A natural idea to avoid the rippling behaviour of FISTA is proposed in [4] by O'Donoghue and Candes. It consists in restarting Nesterov's momentum whenever a non-monotone step is detected in the $f(\beta_k)$ sequence. This non-monotonicity is likely to have been caused by an overshooting of momentum at that step, driving the algorithm out of the optimal direction. To this strategy, Ito *et al.* add in [5] a heuristic, named maintaining top-speed, which avoids restarting the momentum term near the optimum so that its speed up advantage is not lost.

2.4 Decreasing and Stability for L_k

While the convergence proofs in [3] require non-decreasing L_k values, it would also be advantageous for them to be smaller so that we have larger gradient steps whenever possible. In [6] a modification of the original FISTA method is proposed to allow L_k to decrease by diminishing at each iteration the starting L value for backtracking as $\rho_k L_{k-1}$ for some $\rho_k < 1$. This requires to recompute w_k (and the corresponding gradient) and also to adjust the t_k at each step so that they still verify $t_k/L_k \geq t_{k+1}(t_{k+1} - 1)/L_{k+1}$, $\forall k \geq 1$, which is enough for the faster convergence proofs. In [5] the authors also add a practical correction, called stability, which progressively augments the decreasing factor as a trade-off between the decreasing and the backtracking strategies.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Table 1: Dataset sizes and dimensions.

Dataset	Num. Patterns	Dimensions
`year`	46 215	90
`ctscan`	53 500	385
(duke) `breast_cancer`	44	7129
`cpusmall`	6143	12
`leukemia`	72	7129
`ree`	5698	15 960

3 Experiments

We will compare the different strategies above on the 6 regression and binary classification datasets of Table 1. All of them except the `ree` dataset come from the LIBSVM repository; we deal with the classification datasets as regression problems with targets the class labels $\{-1, 1\}$. The goal for the `ree` dataset is to predict wind energy production (kindly provided by Red Eléctrica de España) from numerical weather predictions. Most of the datasets have quite large dimensions, whereas the number of patterns varies from small ones (`leukemia` and `breast_cancer`, which have very large dimensions) to large problems.

We will consider four approaches: starting with standard FISTA and its backtracking extension (FISTA_B), we then group the acceleration techniques of Section 2 into two procedures, FISTA_BR, which adds to FISTA_B Nesterov restarting plus maintaining top speed, and FISTA_BRD, that adds decreasing and stability to the estimation of the Lipschitz constant. The comparison will be done over values of the regularization parameter λ in an equispaced logarithmic path from 10^5 (where all model coefficients will essentially be zero) to 10^{-7} (i.e., essentially no regularization takes place). In order to compare the different methods we first compute for each λ a global optimum value $f^\star$ by running the four methods for 50 000 iterations and then retaining the smallest objective value among the four resulting optima. We then compute for each method and each λ value the number M of iterations needed so that the corresponding $f(\beta_M)$ value verify $\frac{f(\beta_M)-f^\star}{f^\star} \leq 10^{-6}$ (with a limit of 50 000 iterations), and then add for each method these M values over the entire λ path.

Table 2 shows the results, both as total number of iterations and as ratio to the smallest such number. Moreover, Figure 1 depicts the number of iterations for the different λ values. As we can see in both the table and the figure, FISTA_BRD requires the fewest iterations for all problems but `ctscan`, where FISTA_B is better for the smallest λ values and quite close for the rest. For the other datasets, FISTA_BRD is the best across all λ values, particularly, for the optimal regularization parameter obtained by 10-fold cross-validation (vertical dashed line in the plot). Notice that the restarting strategy is not always helpful on its own, since it focuses on avoiding non-monotonicity which, while sensible, is not guaranteed to result in less iterations. On the other hand, the larger step sizes provided by the decreasing strategy clearly help to achieve faster convergence. It seems that the accelerations have larger effects when $d \gg N$ (the case

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Table 2: Iteration results for all the methods and datasets.

Dataset	FISTA		FISTA_{B}		FISTA_{BR}		$\text{FISTA}_{\text{BRD}}$	
	Total	Ratio	Total	Ratio	Total	Ratio	Total	Ratio
`year`	1546	2.55	1104	1.82	779	1.28	607	1.0
`ctscan`	9638	1.22	7895	1.0	12 360	1.57	11 110	1.41
`breast_cancer`	275 976	1.42	271 750	1.40	283 405	1.46	194 301	1.0
`cpusmall`	371	1.78	295	1.41	296	1.42	209	1.0
`leukemia`	189 767	1.40	184 947	1.36	199 686	1.47	135 772	1.0
`ree`	280 718	1.23	273 319	1.20	265 643	1.17	227 737	1.0
Ranking	3.38		2.46		2.92		1.25	

of `leukemia`, `breast_cancer` and, to a smaller degree, `ree`). Also, it appears that the optimal λ somehow balances the problem, so for larger values the algorithms focus on the much easier problem of minimizing the regularization term and, hence, there is less scope for the accelerations to be effective.

4 Discussion and Further Work

Due mainly to its non-monotone behaviour, FISTA is not currently regarded as a state-of-the-art method for solving the Lasso problem. In this paper we have studied the impact of some recent acceleration strategies on the performance of FISTA for regression tasks that can make it much more competitive. Our results lead us to the following conclusions. First, when considering the full regularization path for Lasso there is in most cases a significant benefit from the accelerations studied. Second, there is also a clear advantage when considering only the optimal λ parameter, which is important for production models. Third, a greater gain is obtained for high-dimensional problems, whereas for problems where the number of patterns is much bigger than the dimension the benefits may be not so marked. Finally, as expected the greatest gains are observed for the smallest values of λ, i.e., when the optimization problem is harder. This is clearly the case of `year`, `ctscan` and `cpusmall`, and also of `leukemia`, `breast_cancer` and `ree` (although in these cases we have limited the maximum number of iterations to 50 000).

There are several possible lines for further work along the previous ideas: (i) to take into account the extra costs (function and gradient evaluations) generated by the decreasing strategy, (ii) to consider the effects of the strategies above when warm starts are used while exploring the regularization path, (iii) to compare the iteration performance of FISTA against GLMNet and other competitors and, if still competitive, (iv) to compare the execution times over all the different scenarios to have a better measure of the acceleration effects.

References

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

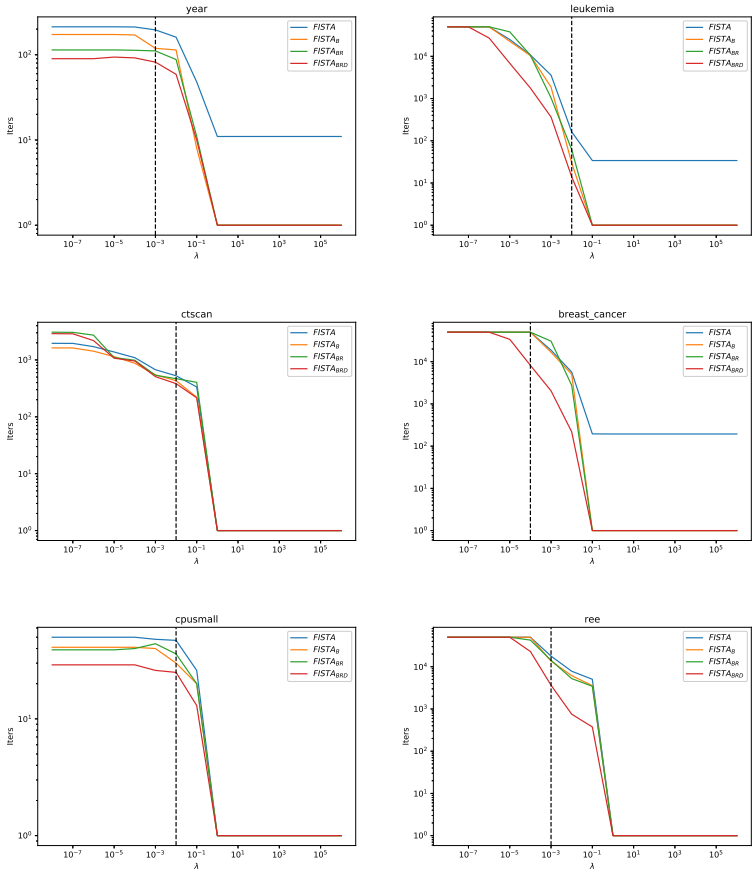

Fig. 1: Iterations until convergence for the full regularization path of Lasso.

[2] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, Jan 2009.

[4] Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, Jul 2013.

[5] Naoki Ito, Akiko Takeda, and Kim-Chuan Toh. A unified formulation and fast accelerated proximal gradient method for classification. *Journal of Machine Learning Research*, 18(16):1–49, 2017.

[6] Katya Scheinberg, Donald Goldfarb, and Xi Bai. Fast first-order methods for composite convex optimization with backtracking. *Found. Comput. Math.*, 14(3):389–417, June 2014.