Convergence rates of nested accelerated inexact proximal methods

Silvia Villa
LCSL, MIT and IIT (Genova, Italy)
silvia.villa@iit.it

Saverio Salzo
DIBRIS, Università di Genova, Italy
saverio.salzo@unige.it

Luca Baldassarre[*]
LIONS, EPFL, Switzerland
luca.baldassarre@epfl.ch

Alessandro Verri
DIBRIS, Università di Genova, Italy
alessandro.verri@unige.it

Abstract

Proximal gradient methods are popular first order algorithms currently used to solve several machine learning and inverse problems. We consider the case where the proximity operator is not available in closed form and is thus approximated via an iterative procedure leading to a nested algorithm. For the first time, we show that relying on an appropriate notion of approximations, which gives an explicit stopping rule for the inner loop, convergence rates for the two-loops algorithm can be proved for accelerated procedures for a large class of approximation algorithms. An experimental comparison with a benchmark primal-dual algorithm is reported and confirms a good empirical performance.

1 Introduction

Accelerated proximal gradient methods [9, 3, 12] are among the most popular first-order techniques to optimize convex composite functionals defined on $\mathbb{R}^d$ of the form

$$F(x) = f(x) + g(x), \qquad\qquad g(x) = \omega(Bx),$$

where f is continuously differentiable with $1/\lambda$ Lipschitz continuous gradient, $B : \mathbb{R}^d \to \mathbb{R}^m$ is a bounded and linear operator, and $\omega : \mathbb{R}^m \to \overline{\mathbb{R}}$ a positively homogeneous convex function, in general nonsmooth. Typical sparsity inducing regularizers in machine learning have this structure (see e.g. [1, 8]). For arbitrary initialization $y_0 = x_0 \in \mathbb{R}^d$, the generic iteration of a well-known instance of such methods, FISTA (fast iterative shrinkage thresholding algorithm), can be written as

$$x_{k+1} = \operatorname{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k)); \qquad y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k) \tag{1}$$

for a sequence of t_k's satisfying $t_0 = 1$ and $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$. Given $\lambda > 0$, $\operatorname{prox}_{\lambda g}$ is the proximity operator of λg defined as [7]

$$\operatorname{prox}_{\lambda g}(y) = \operatorname*{argmin}_{x \in \mathbb{R}^d} \{g(x) + \frac{1}{2\lambda}\|x - y\|^2\}. \tag{2}$$

For several functions g, which are relevant in the applications, the proximity operator is not available in closed form, and the solution of (2) is usually computed via an iterative minimization algorithm. Therefore, the resulting scheme is a two-loops algorithm, constituted of an outer iteration of type (1), and an internal one which serves to compute the prox. The contribution of this paper is the study

[*] LB's research is partly funded by MIRG-268398.

of the global convergence rate of the resulting nested algorithm. We prove for the first time bounds on the convergence rate of the entire two-loops procedure for a variety of inner algorithms. This is done in three steps: first, by introducing a suitable approximation notion for proximal points (see [10]), leading to explicit stopping rules for the inner iterations; second, by proving convergence rates for the inexact implementations of algorithm (1) without taking into account the cost of the computation of the proximity operator (similarly to what has been done in [11]), and finally by adding the costs of the inner loop. The paper is organized accordingly: in Section 2 we introduce admissible approximations of the proximal point, and we state the convergence of the inexact accelerated proximal point algorithm. Moreover, we show a useful characterization in terms of duality gap. In Section 3 we study the global convergence rate of nested iterations. Our results show that for inner algorithms converging sufficiently fast, the global convergence rate is close to the one of the exact scheme. Finally, in Section 4, we show a comparison of the proposed method with a benchmark first order primal-dual algorithm proposed in [4]. Due to space limitation the proofs are not included and are contained in a longer version of the paper [14].

2 Convergence of inexact FISTA

We consider a notion of approximation of proximal points that has been introduced in [5] and recently studied in [10]. It is based on the relaxation of the first order conditions satisfied by exact proximal points, and involves the definition of ϵ-subdifferential.

Definition 1. *Let $\epsilon \geq 0$. We say that $z \in \mathbb{R}^d$ is an* approximation of $\operatorname{prox}_{\lambda g}(y)$ with ϵ-precision *and we write $z \approxeq_\epsilon \operatorname{prox}_{\lambda g}(y)$ if and only if $\dfrac{y-z}{\lambda} \in \partial_{\frac{\epsilon^2}{2\lambda}} g(z)$.*

The approximation requirement is more restrictive than the one proposed in [11], and lead to the following convergence rates for the inexact version of algorithm (1), which show a better dependence on the errors w.r.t. those obtained in [11].

Theorem 1. *Consider algorithm (1), where $x_{k+1} \approxeq_{\epsilon_k} \operatorname{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$. Then, if $\epsilon_k = O(1/k^q)$ with $q > 1/2$, the sequence $(x_k)_{k\in\mathbb{N}}$ is minimizing for F and if the infimum of F is attained the following bounds on the rate of convergence hold true*

$$F(x_k) - \min F = \begin{cases} O\left(1/k^2\right) & \text{if } q > 3/2 \\ O\left(1/k^2\right) + O\left(\log k/k^2\right) & \text{if } q = 3/2 \\ O\left(1/k^2\right) + O\left(1/k^{2q-1}\right) & \text{if } q < 3/2. \end{cases}$$

2.1 Computing admissible approximations

Admissible approximations in the sense of Definition 1 can be characterized in terms of a suitably defined duality gap, and this leads to a very natural test for assessing admissible approximations. The Fenchel-Rockafellar duality formula guarantees that, if δ_K is the indicator function of $K := \partial\omega(0)$ (where ∂ denotes the subdifferential),

$$\Psi_\lambda(v) = \frac{1}{2\lambda}\|\lambda B^T v - y\|^2 + \delta_K(v) - \frac{1}{2\lambda}\|y\|^2\,, \tag{3}$$

is the dual function of the minimization problem of $\Phi_\lambda(x) = g(x) + \frac{1}{2\lambda}\|x-y\|^2$ defining the proximity operator. The duality gap is $G(x,v) := \Phi_\lambda(x) + \Psi_\lambda(v)$ and satisfies $\min_{(x,v)\in\mathbb{R}^d\times\mathbb{R}^m} G(x,v) = 0$. Moreover, if $\bar{v}$ is a solution of the dual problem $\min_v \Psi_\lambda(v)$, then $\bar{z} = y - \lambda B^T\bar{v}$ solves the primal problem (2). This also means that $\min_v G(y - \lambda B^T v, v) = 0$.

Proposition 1. *Let $v \in \mathbb{R}^m$. The following statements are equivalent*

a) $G(y - \lambda B^T v, v) \leq \epsilon^2/(2\lambda)$

c) $\lambda B^T v \approxeq_\epsilon P_{\lambda K}(v)$, *with $P_{\lambda K}$ the proximity operator of $\delta_{\lambda K}$, i.e. the projection onto λK*

b) $y - \lambda B^T v \approxeq_\epsilon \operatorname{prox}_{\lambda g}(y)$.

Proposition 1 shows that admissible approximations can be found by minimizing the duality gap. Moreover, it can be shown that in order to minimize G, it is enough to minimize the dual function Ψ_λ. Thus, admissible approximations can be found for instance by applying FISTA to the dual function Ψ_λ, which is a constrained smooth problem.

3 Asymptotic global iteration complexity

Theorem 1 is mostly of theoretical interest, since it does not take into account the costs due to the computation of the proximal point at each step. Indeed, each iteration of the inexact version of FISTA consists of a gradient descent step, to which we refer to as *external iteration*, and an inner loop, to approximate the proximity operator of g up to a precision ϵ_k. More generally, it can be shown that given an (internal) algorithm that compute x_{k+1} in at most $(D\lambda)/(\epsilon_k^{2/p})$, iterations, with $p > 0$ and $\epsilon_k = O(1/k^q)$, we can bound the global complexity $\mathcal{C}_g$ of the two loops algorithm by

$$\mathcal{C}_g = c_i N_i + c_e N_e = \begin{cases} O\big(1/\epsilon^{\frac{2q/p+1}{2q-1}}\big) + O\big(1/\epsilon^{\frac{1}{2q-1}}\big) & \text{if } 1/2 < q < 3/2 \\ O\big(1/\epsilon^{\frac{2q/p+1}{2}}\big) + O\big(1/\epsilon^{\frac{1}{2}}\big) & \text{if } q > 3/2\,. \end{cases} \tag{4}$$

where c_i and c_e denotes the unitary costs of each type of iteration and N_i and N_e are the total number of inner and outer iterations, respectively. From the estimates above, one can easily see that, in each case, the lower global complexity is reached for $q \to 3/2$ and it is $\mathcal{C}_g = O(1/\epsilon^{\frac{p+3}{2p}+\delta})$ for whatever small $\delta > 0$. Note that, for $p \to +\infty$ we have a complexity of $O(1/\epsilon^{1/2+\delta})$: in other words the global convergence rate tends to $1/N^2$, in the total number N of iterations, and the algorithm behaves once more as an accelerated method. On the other hand, using FISTA to solve the dual problem (3) at each step, gives a theoretical rate of $O(1/\epsilon^{2+\delta})$ (for $q \to 3/2$). Similarly, if we consider an inner algorithm converging linearly, for $q > 3/2$, the resulting convergence rate is $O((1/\epsilon^{1/2})\log(1/\epsilon))$, and thus inexact FISTA is again an accelerated method. We remark that the analysis of the global complexity given above is valid only asymptotically, since we did not estimate any of the constants hidden in the O symbols (in particular c_i and c_e). However, in real situations constants do matter and, in practice, the most effective accuracy rate q is problem dependent and might be different from $3/2$, as we illustrate in the experiments in the next section. The asymptotic point of view also distinguishes our analysis from that in [6], where an accuracy is a priori fixed, and a constant number of internal iterations at each step is shown to be the "optimal" strategy.

4 Numerical Experiments

In this section, we measure the performance of the two loops algorithm inexact FISTA combined with FISTA applied to (3), in comparison with the non accelerated version ISTA combined with FISTA applied to (3), and with the primal-dual algorithm proposed in [4] (PRIDU). Following Theorem 1, we consider sequences of errors of type $\epsilon_k = C/k^q$, with q, hereafter referred as *accuracy rate*, chosen between 0.1 and 1.7. We analyze two well-known problems: deblurring with total variation regularization and learning a linear estimator via regularized empirical risk minimization with the overlapping group lasso penalty (OGL problem). When taking into account the cost of computing the proximity operator, there is a trade-off between the number of external and internal iterations. Since internal and external iterations in general have different computational costs — which depend on the specific problem considered and the machine CPU — the total number of iterations is not a good measure of the algorithm's performance. Therefore, for all algorithms we provide the number of external and internal iterations and the CPU time needed to reach a desired accuracy for the relative difference to the optimal value. We use the *warm-restart* procedure, consisting in initializing the internal algorithm with the solution obtained at the previous step. We empirically observed that this initialization strategy drastically reduces the total number of iterations and speeds up the algorithm. All the numerical experiments have been performed in MATLAB, on an iMac with Intel Core i5 CPU, 2,5 Ghz, 6MB cache L3, and 6 GB of RAM. For the deblurring problem, we followed the same experimental setup as in [2]. The OGL problem has been generated from the breast cancer dataset provided by [13]. The structure of the overlapping groups gives rise to a matrix B of size 15126×3510. Despite the high dimensionality, one can take advantage of its sparseness. We analyze the choice of the regularization parameter $\tau = 0.01$. As concerns the TV problem, inexact FISTA+FISTA ($q = 1.3$ or $q = 1.5$) outperforms both PRIDU and ISTA, for high precisions. PRIDU exhibits a fast convergence at the beginning, but then slows down for higher precisions. For the OGL problem, and precision 10^{-4}, inexact FISTA is the fastest. For the middle precision, the algorithms' performances are comparable. For the highest precision, PRIDU and ISTA perform better. Summarizing, the performance of Algorithm 1 combined with FISTA on the dual and warm restart is comparable with state-of-the-art algorithms, being sometimes better. To

Table 1: **Deblurring with Total Variation regularization (top) and Breast cancer dataset: Overlapping Group Lasso (bottom)**. Performance evaluation of inexact FISTA, ISTA and PRIDU, corresponding to different choices of the parameters q, and σ suggested by the authors, respectively. Concerning inexact FISTA and ISTA, the results are reported only for the q's giving the best results. The entries in the table refer to the CPU time (in seconds) needed to reach a relative difference w.r.t. to the optimal value below the thresholds 10^{-4}, 10^{-6} and 10^{-8}, the number of external iterations (# Ext), and the total number of internal interations (# Int).

Precision	10^{-4}			10^{-6}			10^{-8}		
Algo	Time	# Ext	# Int	Time	# Ext	# Int	Time	# Ext	# Int
in FISTA $q = 1.3$	16.2	118	1600	**63.6**	387	6437	**272.1**	1300	28350
ISTA $q = 0.1$	36.9	1341	1341	147.2	5346	5346	635.4	23031	23031
PRIDU $\sigma = 10$	**7.4**	362	-	165.7	8186	-	4684	231848	-

Precision	10^{-4}			10^{-6}			10^{-8}		
Algo	Time	# Ext	# Int	Time	# Ext	# Int	Time	# Ext	# Int
in FISTA $q = 1.3$	**2.1**	51	2103	11.2	247	11389	60.4	1179	61915
ISTA $q = 0.5$	4.4	1217	1827	**9.5**	2850	3460	**14.9**	4603	5213
PRIDU $\sigma = 1.07$	5.8	1602	-	11.0	3026	-	16.1	4452	-

this purpose, the experiments also give some guidelines for choosing the parameter q. We also show situations where the acceleration is lost, in particular referring to high precision.

References

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[2] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring. *IEEE Trans. Image Proc.*, 18(11):2419–2434, 2009.

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

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[5] B. Lemaire. About the convergence of the proximal method. In *Advances in optimization (Lambrecht, 1991)*, volume 382 of *Lecture Notes in Econom. and Math. Systems*, pages 39–51. Springer, Berlin, 1992.

[6] P. Machart, S. Anthoine and L. Baldassarre. Optimal computational trade-off of inexact proximal methods `http://arxiv.org/abs/1210.5034`,2012

[7] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[8] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. In *Machine Learning and Knowledge Discovery in Databases*, Springer, 2010.

[9] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, 2005.

[10] S. Salzo and S. Villa. Inexact and accelerated proximal point algorithm. *J. Convex Anal.*, 19(4), 2012.

[11] M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. arXiv:1109.2415v2.

[12] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125:263–295, 2010. 10.1007/s10107-010-0394-2.

[13] M.J. Van De Vijver et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[14] S. Villa, S. Salzo, L. Baldassarre and A. Verri. Accelerated and inexact forward-backward algorithms. `http://www.optimization-online.org/DB_HTML/2011/08/3132.html`,2011