Local Linear Convergence of Inertial Forward-Backward Splitting for Low Complexity Regularization

Jingwei Liang, Jalal M. Fadili, Gabriel Peyré

Local Linear Convergence of Inertial Forward–Backward Splitting for Low Complexity Regularization

Jingwei Liang and Jalal M. Fadili
CNRS, GREYC, ENSICAEN, Université de Caen
Email: {Jingwei.Liang, Jalal.Fadili}@ensicaen.fr

Gabriel Peyré
CNRS, Ceremade, Université Paris-Dauphine
Email: Gabriel.Peyre@ceremade.dauphine.fr

***Abstract*—In this abstract, we consider the inertial Forward-Backward (iFB) splitting method and its special cases (Forward–Backward/ISTA and FISTA). Under the assumption that the non-smooth part of the objective is *partly smooth* relative to an active smooth manifold, we show that iFB-type methods (i) identify the active manifold in finite time, then (ii) enter a local linear convergence regime that we characterize precisely. This gives a grounded and unified explanation to the typical behaviour that has been observed numerically for many low-complexity regularizers, including ℓ_1, $\ell_{1,2}$-norms, total variation (TV) and nuclear norm to name a few. The obtained results are illustrated by concrete examples.**

I. Introduction

Consider the following structured optimization problem

$$\min_{x\in\mathbb{R}^n} \big\{\Phi(x) \stackrel{\text{def}}{=} F(x) + J(x)\big\}, \qquad (\mathcal{P})$$

where $J \in \Gamma_0(\mathbb{R}^n)$, the set of proper, lower semi-continuous and convex functions, F is convex, $C^{1,1}(\mathbb{R}^n)$ with ∇F being β-Lipschitz continuous. We assume that $\operatorname{Argmin}\Phi \neq \emptyset$.

In this paper, we consider a generic form of inertial Forward–Backward for solving $(\mathcal{P})$ which reads,

$$\begin{aligned} y_a^k = x^k + a_k(x^k - x^{k-1}), \quad y_b^k = x^k + b_k(x^k - x^{k-1}),\\ x^{k+1} = \operatorname{Prox}_{\gamma_k J}\big(y_a^k - \gamma_k \nabla F(y_b^k)\big), \end{aligned} \qquad (\text{I.1})$$

where $a_k \in [0, \bar{a}]$ and $b_k \in [0, \bar{b}]$, $(\bar{a}, \bar{b}) \in [0,1]^2$, and the step-size $0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2/\beta$. For $\gamma > 0$, the proximity operator is defined as $\operatorname{Prox}_{\gamma J}(x) = \operatorname{argmin}_{z\in\mathbb{R}^n} \frac{1}{2\gamma}\|z-x\|^2 + J(z)$.

iFB (I.1) covers various special cases in the literature, including the (unrelaxed) Forward–Backward (FB) [1] and FISTA [2]. In the original FISTA, only convergence of the objective function is guaranteed. Recently in [5], the iterates are proved to be convergent under $a_k = b_k = (t_{k-1} - 1)/t_k$ where $t_k = (k+p-1)/p, p \ge 2$.

II. Partly Smooth Functions and Finite Identification

The class of partly smooth functions [3], is specialized here to functions in $\Gamma_0(\mathbb{R}^n)$. Denote $\operatorname{par}(C)$ the linear subspace parallel to the non-empty convex set $C \subset \mathbb{R}^n$, and $\operatorname{ri}(C)$ its relative interior.

Definition II.1. Let $J \in \Gamma_0(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$ such that $\partial J(x) \neq \emptyset$. J is *partly smooth at x relative to a set $\mathcal{M}$* containing x if

- (Smoothness) $\mathcal{M}$ is a C^2-manifold, $J|_{\mathcal{M}}$ is C^2 around x;
- (Sharpness) The tangent space $\mathcal{T}_{\mathcal{M}}(x) = T_x \stackrel{\text{def}}{=} \operatorname{par}(\partial J(x))^\perp$;
- (Continuity) The ∂J is continuous at x relative to $\mathcal{M}$.

Examples of such functions are given in Section IV, see also [4].

Theorem II.2 (Finite activity identification). *Suppose x^k converges to a minimizer $x^\star$ of $(\mathcal{P})$ such that J is partly smooth at $x^\star$ relative to $\mathcal{M}_{x^\star}$, and*

$$-\nabla F(x^\star) \in \operatorname{ri}\big(\partial J(x^\star)\big), \qquad (\text{II.1})$$

then there exists a $K > 0$ such that for all $k \ge K$, $x^k \in \mathcal{M}_{x^\star}$. If moreover $\mathcal{M}_{x^\star}$ is affine/linear, then $y_a^k, y_b^k \in \mathcal{M}_{x^\star}$ for $k > K$.

Condition (II.1) can be viewed as a geometric generalization of the strict complementarity of non-linear programming, and is almost necessary for the finite identification of $\mathcal{M}_{x^\star}$ [3].

III. Local Linear Convergence

We now turn to the local linear convergence of the iFB-type methods with partly smooth functions. For space limitations, we mainly focus on the case where $a_k = b_k$, and denote $d^{k+1} = \begin{pmatrix} x^{k+1} - x^\star \\ x^k - x^\star \end{pmatrix}$.

Theorem III.1. *We assume the conditions of Theorem II.2 hold. If moreover F is C^2 near $x^\star$ and there exists $\alpha \ge 0$ such that $\mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}} \succ \alpha \mathrm{Id}$. Then for all k large enough, we have*

1) Q-linear rate*: if $0 < \underline{\gamma} \le \gamma_k \le \bar{\gamma} < \min(2\alpha\beta^{-2}, 2\beta^{-1})$, then given any ρ such that $1 > \rho \ge \widetilde{\rho}_k$, the iterates satisfy*

$$\|x^{k+1} - x^\star\|^2 \le \|d^{k+1}\|^2 \le \rho \|d^k\|^2,$$

where $\eta = \max\big\{q(\underline{\gamma}), q(\bar{\gamma})\big\} \in [0,1[$, $q(\gamma) = 1 - 2\alpha\gamma + \beta^2\gamma^2$,

$$\widetilde{\rho}_k = \begin{cases} \frac{(1+a_k)\eta + \sqrt{(1+a_k)^2\eta^2 - 4a_k\eta}}{2}, & \eta \in [\frac{4a_k}{(1+a_k)^2}, 1[, \\ \sqrt{a_k\eta}, & \eta \in [0, \frac{4a_k}{(1+a_k)^2}]. \end{cases}$$

2) R-linear rate*: if $\mathcal{M}_{x^\star}$ is affine/linear, then*

$$\|x^{k+1} - x^\star\|^2 \le \|d^{k+1}\|^2 \le \rho_k \|d^k\|^2,$$

where $\rho_k \in [0,1[$

$$\rho_k = \begin{cases} \frac{|(1+a_k)\eta_k| + \sqrt{(1+a_k)^2\eta_k^2 - 4a_k\eta_k}}{2}, & \eta_k \in]-1,0] \cup [\frac{4a_k}{(1+a_k)^2}, 1[, \\ \sqrt{a_k\eta_k}, & \eta_k \in [0, \frac{4a_k}{(1+a_k)^2}], \end{cases}$$

and $\eta_k \in]-1,1[$ is an eigenvalue of $\mathrm{Id} - \gamma_k \mathrm{P}_T \int_0^1 \nabla^2 F(x^\star + t(y_a^k - x^\star))dt \mathrm{P}_T$.

IV. Numerical Experiments

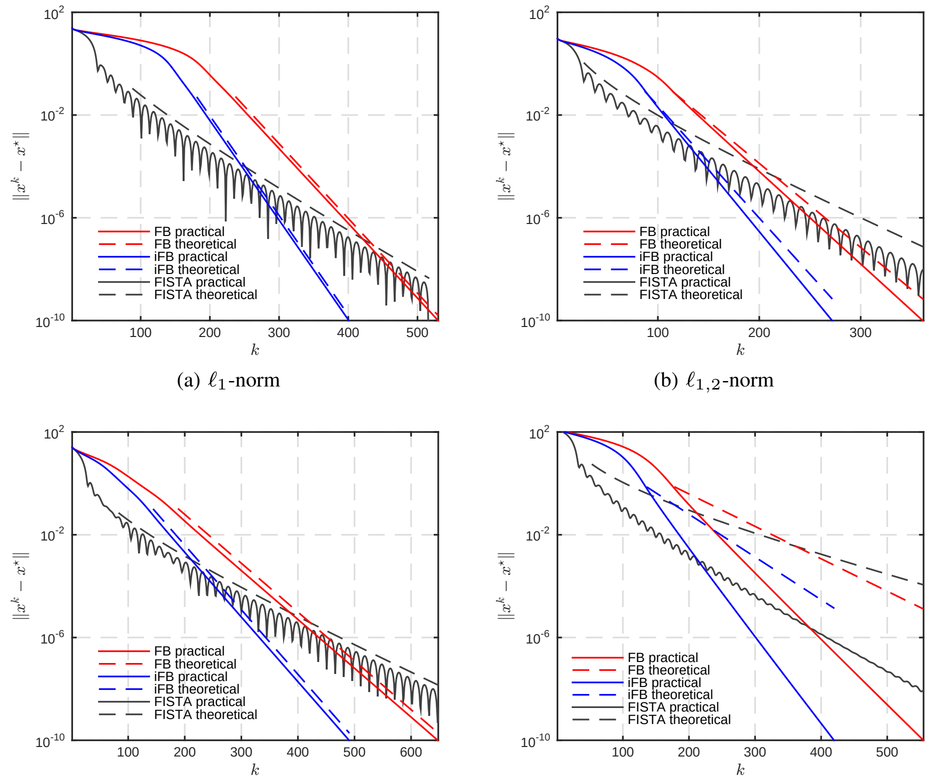
Example IV.1 (ℓ_1-norm). The ℓ_1-norm is partly smooth relative to $\mathcal{M} = T_x = \{u \in \mathbb{R}^n : \operatorname{supp}(u) \subseteq \operatorname{supp}(x)\}$.

Example IV.2 ($\ell_{1,2}$-norm). $\ell_{1,2}$-norm is partly smooth relative to $\mathcal{M} = T_x = \{u \in \mathbb{R}^n : \operatorname{supp}_{\mathcal{B}}(u) \subseteq \operatorname{supp}_{\mathcal{B}}(x)\}$, where $\operatorname{supp}_{\mathcal{B}}(x) = \bigcup\{b : x_b \neq 0\}$, and $\bigcup_{b\in\mathcal{B}} b = \{1, \ldots, n\}$.

Example IV.3 (TV semi-norm). The TV semi-norm $\|x\|_{\mathrm{TV}} = \|\nabla x\|_1$ is partly smooth relative the subspace $\mathcal{M} = T_x = \big\{u \in \mathbb{R}^n : \operatorname{supp}(\nabla u) \subseteq I\big\}$, $I = \operatorname{supp}(\nabla x)$.

Example IV.4 (Nuclear norm). The nuclear norm is partly smooth relative to the manifold of fixed rank matrices, $\mathcal{M} = \big\{z \in \mathbb{R}^{n_1\times n_2} : \operatorname{rank}(z) = r\big\}$.

We now consider the problem $\min_{x\in\mathbb{R}^n} \frac{1}{2}\|y - Ax\|^2 + \lambda J(x)$, where $y \in \mathbb{R}^m$ is the observation, $A : \mathbb{R}^n \to \mathbb{R}^m$ is drawn from the standard Gaussian ensemble, and $\lambda > 0$ is the regularization parameter. The convergence profiles are depicted in Figure 1.



(a) ℓ_1-norm (b) $\ell_{1,2}$-norm

(c) TV semi-norm (d) Nuclear norm

Fig. 1: Local linear convergence of iFB-type methods in terms of $\|x^k - x^\star\|$. The forward model of the problem of interests reads $y = Ax_0 + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \delta^2)$. (a) ℓ_1-norm, $(m, n) = (48, 128)$, x_0 is 8-sparse; (b) $\ell_{1,2}$-norm, $(m, n) = (60, 128)$, x_0 has 3 non-zero blocks with block-size 4; (c) 1D TV semi-norm, $(m, n) = (48, 128)$, ∇x_0 is 8-sparse; (d) Nuclear norm, $(m, n) = (1425, 2500)$, $x_0 \in \mathbb{R}^{50\times 50}$ and $\mathrm{rank}(x_0) = 5$. The *red*, *black* and *blue* lines are respectively the results of FB, FISTA [5] and iFB (with $a_k = b_k \equiv \sqrt{5} - 2.01$). All algorithms were tested with $\gamma_k \equiv 1/\|A\|^2$. The *solid* lines are the practical observed profiles and the *dashed* ones the theoretical predictions. The beginning of the dashed lines are the points when x^k identifies the manifold $\mathcal{M}_{x^\star}$. As one can observe, FISTA has the fastest manifold identification, however, locally it is the slowest for all tested examples. Indeed, when the manifold is affine, it can be shown from Theorem III.1 that $\rho_k \in]\eta_k, \sqrt{\eta_k}]$ for $a_k > \eta_k$, *i.e.* FISTA is locally slower than FB.

Acknowledgment

This work has been partly supported by the European Research Council (ERC project SIGMA-Vision). JF was partly supported by Institut Universitaire de France.

References

[1] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16(6):964–979, 1979.
[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.
[3] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13(3):702–725, 2003.
[4] S. Vaiter, G. Peyré, and M. J. Fadili, *Partly smooth regularization of inverse problems*, arXiv:1405.1004, 2014.
[5] A. Chambolle and D. Dossal, *How to make sure the iterates of FISTA converge*, 2014.