The Rate of Convergence of Nesterov's Accelerated Forward-Backward Method is Actually Faster Than $1/k^2$ — Source link

Hedy Attouch, Juan Peypouquet

Institutions: Valparaiso University

Published on: 01 Sep 2016 - Siam Journal on Optimization (Society for Industrial and Applied Mathematics)

Topics: Separable space and Rate of convergence

THE RATE OF CONVERGENCE OF NESTEROV'S ACCELERATED FORWARD-BACKWARD METHOD IS ACTUALLY $o(k^{-2})$

HEDY ATTOUCH AND JUAN PEYPOUQUET

Abstract. The *forward-backward algorithm* is a powerful tool for solving optimization problems with a *additively separable* and *smooth* + *nonsmooth* structure. In the convex setting, a simple but ingenious acceleration scheme developed by Nesterov has been proved useful to improve the theoretical rate of convergence for the function values from the standard $\mathcal{O}(k^{-1})$ down to $\mathcal{O}(k^{-2})$. In this short paper, we prove that the rate of convergence of a slight variant of Nesterov's accelerated forward-backward method, which produces *convergent* sequences, is actually $o(k^{-2})$, rather than $\mathcal{O}(k^{-2})$. Our arguments rely on the connection between this algorithm and a second-order differential inclusion with vanishing damping.

Introduction

Let $\mathcal{H}$ be a real Hilbert space endowed with the scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, and consider the problem

$$\min \{ \Psi(x) + \Phi(x) : \ x \in \mathcal{H} \} \tag{1}$$

where $\Psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper lower-semicontinuous convex function, and $\Phi : \mathcal{H} \to \mathbb{R}$ is a continuously differentiable convex function, whose gradient is Lipschitz continuous.

Based on the *gradient projection* algorithm of [9] and [10], the *forward-backward* method was proposed in [11], and [20] to overcome the inherent difficulties of minimizing the nonsmooth sum of two functions, as in (1), while exploiting its *additively separable* and *smooth* + *nonsmooth* structure. It gained popularity in image processing following [8] and [7]: when Ψ is the ℓ^1 norm in $\mathbb{R}^N$ and Φ is quadratic, this gives the *Iterative Shrinkage-Thesholding Algorithm* (ISTA). Some time later, a decisive improvement came with [4], where ISTA was successfully combined with Nesterov's acceleration scheme [14] producing the *Fast Iterative Shrinkage-Thesholding Algorithm* (FISTA). For general Φ and Ψ, and after some simplification, the *Accelerated Forward-Backward* method can be written as

$$\left\{ \begin{array}{rcl} y_k & = & x_k + \frac{k-1}{k+\alpha-1}(x_k - x_{k-1}) \\ \\ x_{k+1} & = & \operatorname{prox}_{s\Psi}\left(y_k - s(\nabla\Phi(y_k))\right), \end{array} \right. \tag{2}$$

where $\alpha > 0$ and $s > 0$. This algorithm is also in close connection with the proximal-based inertial algorithms [1], [13] and [22]. The choice $\alpha = 3$ is current common practice. The remarkable property of this algorithm is that, despite its simplicity and computational efficiency −equivalent to that of the classical forward-backward method−, it guarantees a rate of convergence of $\mathcal{O}(k^{-2})$, where k is the number of iterations, for the minimization of the function values, instead of the classical $\mathcal{O}(k^{-1})$ that is obtained for the unaccelerated counterpart. However, while sequences generated by the classical forward backward method are convergent, the convergence of the sequence (x_k) generated by (2) to a minimizer of $\Phi + \Psi$ puzzled researchers for over two decades. This question was recently settled in [5] and [2] independently, and using different arguments. In [5], the authors use a *descent* inequality satisfied by forward-backward iterations. A perspicuous abstract presentation of this idea is given in [6, Section 2.2]. In turn, the proof given in [2] relies on the connection between (2) and the differential inclusion

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial\Psi(x(t)) + \nabla\Phi(x(t)) \ni 0. \tag{3}$$

Indeed, as pointed out in [25, 2], algorithm (2) can be seen as an appropriate finite-difference discretization of (3). In [25], the authors studied

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\Theta(x(t)) = 0. \tag{4}$$

and proved that

$$\Theta(x(t)) - \min\Theta = \mathcal{O}(t^{-2})$$

Key words and phrases. Convex optimization, fast convergent methods, Nesterov method.

Effort sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-14-1-0056. Also supported by Fondecyt Grant 1140829, Conicyt Anillo ACT-1106, ECOS-Conicyt Project C13E03, Millenium Nucleus ICM/FIC RC130003, Conicyt Project MATHAMSUD 15MATH-02, Conicyt Redes 140183, and Basal Project CMM Universidad de Chile. Part of this research was carried out while the authors were visiting Hangzhou Dianzi University by invitation of Professor Hong-Kun Xu.

when $\alpha \geq 3$. Convergence of the trajectories was obtained in [2] for $\alpha > 3$. The study of the long-term behavior of the trajectories satisfying this evolution equation has given important insight into Nesterov's acceleration method and its variants, and the present work is inspired in this relationship. If $\alpha > 3$, we actually have

$$\Theta(x(t)) - \min \Theta = o(t^{-2}).$$

Although it can be derived from the arguments in [2], it was May [12] who first pointed out this fact, giving a different proof. This is another justification for the interest of taking $\alpha > 3$ instead of $\alpha = 3$.

The purpose of this paper is to show that sequences generated by Nesterov's accelerated version of the forward-backward method approximate the optimal value of the problem with a rate that is strictly faster than $\mathcal{O}(k^{-2})$. More precisely, we prove the following:

Theorem 1. *Let $\Psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper, lower-semicontinuous and convex, and let $\Phi : \mathcal{H} \to \mathbb{R}$ be convex and continuously differentiable with L-Lipschitz continuous gradient. Suppose that $S = \operatorname{argmin}(\Psi + \Phi) \neq \emptyset$, and let (x_k) be a sequence generated by algorithm (2) with $\alpha > 3$ and $0 < s < \frac{1}{L}$. Then, the function values and the velocities satisfy*

$$\lim_{k\to\infty} k^2 \Big((\Psi + \Phi)(x_k) - \min(\Psi + \Phi) \Big) = 0 \qquad \text{and} \qquad \lim_{k\to\infty} k\|x_{k+1} - x_k\| = 0,$$

respectively. In other words,

$$(\Psi + \Phi)(x_k) - \min(\Psi + \Phi) = o(k^{-2}) \qquad \text{and} \qquad \|x_{k+1} - x_k\| = o(k^{-1}).$$

Moreover, we recover some results from [2, Section 5], closely connected with the ones in [5], with simplified arguments. As shown in [2, Example 2.13], there is no $p > 2$ such that the order of convergence is $\mathcal{O}(k^{-p})$ for every Φ and Ψ. In this sense, Theorem 1 is optimal.

We close this paper by establishing a tolerance estimation that guarantees that the order of convergence is preserved when the iterations given in (2) are computed inexactly (see Theorem 4). Inexact FISTA-like algorithms have also been considered in [23, 24].

1. Main results

Throughout this section, $\Psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is proper, lower-semicontinuous and convex, and $\Phi : \mathcal{H} \to \mathbb{R}$ is convex and continuously differentiable with L-Lipschitz continuous gradient. To simplify the notation, we set $\Theta = \Psi + \Phi$. We assume that $S = \operatorname{argmin}(\Psi + \Phi) \neq \emptyset$, and consider a sequence (x_k) generated by algorithm (2) with $\alpha \geq 3$ and $0 < s < \frac{1}{L}$. For standard notation and convex analysis background, see [3, 21].

1.1. **Some important estimations.** We begin by establishing the basic properties of the sequence (x_k). Some results can be found in [5, 2], for which we provide simplified proofs.

Let $x^* \in \operatorname{argmin} \Theta$. For each $k \in \mathbb{N}$, set

$$\mathcal{E}(k) := \frac{2s}{\alpha - 1} (k + \alpha - 2)^2 \left(\Theta(x_k) - \Theta(x^*)\right) + (\alpha - 1)\|z_k - x^*\|^2, \tag{5}$$

where

$$z_k := \frac{k + \alpha - 1}{\alpha - 1} y_k - \frac{k}{\alpha - 1} x_k = x_k + \frac{k - 1}{\alpha - 1}(x_k - x_{k-1}). \tag{6}$$

The key idea is to verify that the sequence $(\mathcal{E}(k))$ has Lyapunov-type properties. By introducing the operator $G_s : \mathcal{H} \to \mathcal{H}$, defined by

$$G_s(y) = \frac{1}{s} \left(y - \operatorname{prox}_{s\Psi} \left(y - s\nabla\Phi(y)\right)\right)$$

for each $y \in \mathcal{H}$, the formula for x_{k+1} in algorithm (2) can be rewritten as

$$x_{k+1} = y_k - sG_s(y_k). \tag{7}$$

The variable z_k, defined in (6), will play an important role. Simple algebraic manipulations give

$$z_{k+1} = \frac{k + \alpha - 1}{\alpha - 1} \left(y_k - sG_s(y_k)\right) - \frac{k}{\alpha - 1} x_k = z_k - \frac{s}{\alpha - 1} (k + \alpha - 1)\, G_s(y_k). \tag{8}$$

The operator G_s satisfies

$$\Theta(y - sG_s(y)) \leq \Theta(x) + \langle G_s(y), y - x\rangle - \frac{s}{2}\|G_s(y)\|^2. \tag{9}$$

for all $x, y \in \mathcal{H}$ (see [4], [5], [19], [25]), since $s \leq \frac{1}{L}$, and $\nabla\Phi$ is L-lipschitz continuous. Let us write successively this formula at $y = y_k$ and $x = x_k$, then at $y = y_k$ and $x = x^*$. We obtain

$$\Theta(y_k - sG_s(y_k)) \leq \Theta(x_k) + \langle G_s(y_k), y_k - x_k\rangle - \frac{s}{2}\|G_s(y_k)\|^2 \tag{10}$$

and

$$\Theta(y_k - sG_s(y_k)) \leq \Theta(x^*) + \langle G_s(y_k), y_k - x^*\rangle - \frac{s}{2}\|G_s(y_k)\|^2, \tag{11}$$

respectively. Multiplying the first inequality by $\frac{k}{k+\alpha-1}$, and the second one by $\frac{\alpha-1}{k+\alpha-1}$, then adding the two resulting inequalities, and using the fact that $x_{k+1} = y_k - sG_s(y_k)$, we obtain

$$\Theta(x_{k+1}) \leq \frac{k}{k+\alpha-1}\Theta(x_k) + \frac{\alpha-1}{k+\alpha-1}\Theta(x^*) - \frac{s}{2}\|G_s(y_k)\|^2 + \left\langle G_s(y_k), \frac{k}{k+\alpha-1}(y_k - x_k) + \frac{\alpha-1}{k+\alpha-1}(y_k - x^*)\right\rangle.$$

Since

$$\frac{k}{k+\alpha-1}(y_k - x_k) + \frac{\alpha-1}{k+\alpha-1}(y_k - x^*) = \frac{\alpha-1}{k+\alpha-1}(z_k - x^*),$$

we obtain

$$\Theta(x_{k+1}) \leq \frac{k}{k+\alpha-1}\Theta(x_k) + \frac{\alpha-1}{k+\alpha-1}\Theta(x^*) + \frac{\alpha-1}{k+\alpha-1}\langle G_s(y_k), z_k - x^*\rangle - \frac{s}{2}\|G_s(y_k)\|^2. \tag{12}$$

We shall obtain a recursion from (12). To this end, observe that (8) gives

$$z_{k+1} - x^* = z_k - x^* - \frac{s}{\alpha-1}(k+\alpha-1)\, G_s(y_k).$$

After developing

$$\|z_{k+1} - x^*\|^2 = \|z_k - x^*\|^2 - 2\frac{s}{\alpha-1}(k+\alpha-1)\langle z_k - x^*, G_s(y_k)\rangle + \frac{s^2}{(\alpha-1)^2}(k+\alpha-1)^2\|G_s(y_k)\|^2,$$

and multiplying the above expression by $\frac{(\alpha-1)^2}{2s(k+\alpha-1)^2}$, we obtain

$$\frac{(\alpha-1)^2}{2s\,(k+\alpha-1)^2}\left(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2\right) = \frac{\alpha-1}{k+\alpha-1}\langle G_s(y_k), z_k - x^*\rangle - \frac{s}{2}\|G_s(y_k)\|^2.$$

Replacing this in (12), we deduce that

$$\Theta(x_{k+1}) \leq \frac{k}{k+\alpha-1}\Theta(x_k) + \frac{\alpha-1}{k+\alpha-1}\Theta(x^*) + \frac{(\alpha-1)^2}{2s\,(k+\alpha-1)^2}\left(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2\right).$$

Equivalently,

$$\Theta(x_{k+1}) - \Theta(x^*) \leq \frac{k}{k+\alpha-1}\left(\Theta(x_k) - \Theta(x^*)\right) + \frac{(\alpha-1)^2}{2s\,(k+\alpha-1)^2}\left(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2\right).$$

Multiplying by $\frac{2s}{\alpha-1}(k+\alpha-1)^2$, we obtain

$$\frac{2s}{\alpha-1}(k+\alpha-1)^2\left(\Theta(x_{k+1}) - \Theta(x^*)\right) \leq \frac{2s}{\alpha-1}k\,(k+\alpha-1)\left(\Theta(x_k) - \Theta(x^*)\right) + (\alpha-1)\left(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2\right),$$

which implies

$$\begin{aligned}\frac{2s}{\alpha-1}(k+\alpha-1)^2\left(\Theta(x_{k+1}) - \Theta(x^*)\right) &+ 2s\frac{\alpha-3}{\alpha-1}k\left(\Theta(x_k) - \Theta(x^*)\right)\\ &\leq \frac{2s}{\alpha-1}(k+\alpha-2)^2\left(\Theta(x_k) - \Theta(x^*)\right) + (\alpha-1)\left(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2\right),\end{aligned}$$

in view of

$$k\,(k+\alpha-1) = (k+\alpha-2)^2 - k(\alpha-3) - (\alpha-2)^2 \leq (k+\alpha-2)^2 - k(\alpha-3).$$

In other words,

$$\mathcal{E}(k+1) + 2s\frac{\alpha-3}{\alpha-1}k\left(\Theta(x_k) - \Theta(x^*)\right) \leq \mathcal{E}(k). \tag{13}$$

We deduce the following:

Fact 1. *The sequence $\big(\mathcal{E}(k)\big)$ is nonincreasing and $\lim_{k\to\infty}\mathcal{E}(k)$ exists.*

In particular, $\mathcal{E}(k) \leq \mathcal{E}(0)$ and we have:

Fact 2. *For each $k \geq 0$, we have $\Theta(x_k) - \Theta(x^*) \leq \frac{(\alpha-1)\mathcal{E}(0)}{2s(k+\alpha-2)^2}$ and $\|z_k - x^*\|^2 \leq \frac{\mathcal{E}(0)}{\alpha-1}$.*

From (13), we also obtain:

Fact 3. *If $\alpha > 3$, then $\sum_{k=1}^{\infty} k\Big(\Theta(x_k) - \Theta(x^*)\Big) \leq \frac{(\alpha-1)\mathcal{E}(1)}{2s(\alpha-3)}$.*

Now, using (10) and recalling that $x_{k+1} = y_k - sG_s(y_k)$ and $y_k - x_k = \frac{k-1}{k+\alpha-1}(x_k - x_{k-1})$, we obtain

$$\Theta(x_{k+1}) + \frac{1}{2s}\|x_{k+1} - x_k\|^2 \leq \Theta(x_k) + \frac{1}{2s}\frac{(k-1)^2}{(k+\alpha-1)^2}\|x_k - x_{k-1}\|^2. \tag{14}$$

Subtract $\Theta(x^*)$ on both sides, and set $\theta_k := \Theta(x_k) - \Theta(x^*)$ and $d_k := \frac{1}{2s}\|x_{k+1} - x_k\|^2$. We can write (14) as

$$\theta_{k+1} + d_k \leq \theta_k + \frac{(k-1)^2}{(k+\alpha-1)^2} d_{k-1}. \tag{15}$$

Since $k + \alpha - 1 \geq k + 1$, (15) implies

$$(k+1)^2 d_k - (k-1)^2 d_{k-1} \leq (k+1)^2(\theta_k - \theta_{k+1}).$$

But then

$$(k+1)^2(\theta_k - \theta_{k+1}) = k^2\theta_k - (k+1)^2\theta_{k+1} + (2k+1)\theta_k \leq k^2\theta_k - (k+1)^2\theta_{k+1} + 3k\theta_k$$

for $k \geq 1$, and so

$$\begin{aligned} 2kd_k + k^2 d_k - (k-1)^2 d_{k-1} &\leq (k+1)^2 d_k - (k-1)^2 d_{k-1} \\ &\leq (k+1)^2(\theta_k - \theta_{k+1}) \\ &\leq k^2\theta_k - (k+1)^2\theta_{k+1} + 3k\theta_k \end{aligned}$$

for $k \geq 1$. Summing for $k = 1, \ldots, K$, we obtain

$$K^2 d_K + 2\sum_{k=1}^{K} k d_k \leq \theta_1 + \frac{3(\alpha-1)\mathcal{E}(1)}{2s(\alpha-3)}$$

in view of Fact 3. In particular, we obtain

Fact 4. *If $\alpha > 3$, then* $\displaystyle\sum_{k=1}^{\infty} k\|x_{k+1} - x_k\|^2 \leq \frac{\alpha(3\alpha-5)\mathcal{E}(1)}{4s(\alpha-1)(\alpha-3)}$.

Remark 1. Observe that the upper bounds given in Facts 3 and 4 tend to ∞ as α tends to 3.

1.2. **From $\mathcal{O}(k^{-2})$ to $o(k^{-2})$.** Recall that $\Psi : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ is proper, lower-semicontinuous and convex, $\Phi : \mathcal{H} \to \mathbb{R}$ is convex and continuously differentiable with L-Lipschitz continuous gradient, and $\Theta = \Phi + \Psi$. We suppose that $S = \operatorname{argmin}(\Psi + \Phi) \neq \emptyset$, and let (x_k) be a sequence generated by algorithm (2) with $\alpha > 3$ and $0 < s < \frac{1}{L}$. We shall prove that the function values and the velocities satisfy

$$\lim_{k\to\infty} k^2\Big((\Psi + \Phi)(x_k) - \min(\Psi + \Phi)\Big) = 0 \qquad \text{and} \qquad \lim_{k\to\infty} k\|x_{k+1} - x_k\| = 0,$$

respectively. In other words, $(\Psi + \Phi)(x_k) - \min(\Psi + \Phi) = o(k^{-2})$ and $\|x_{k+1} - x_k\| = o(k^{-1})$.

The following result is new, and will play a central role in the proof of Theorem 1.

Lemma 2. *If $\alpha > 3$, then* $\displaystyle\lim_{k\to\infty}\Big[k^2\|x_{k+1} - x_k\|^2 + (k+1)^2\big(\Theta(x_{k+1}) - \Theta(x^*)\big)\Big]$ *exists.*

Proof. Since $k + \alpha - 1 \geq k$, inequality (15) gives

$$k^2 d_k - (k-1)^2 d_{k-1} \leq k^2(\theta_k - \theta_{k+1}).$$

But

$$(k+1)^2\theta_{k+1} - k^2\theta_k = k^2(\theta_{k+1} - \theta_k) + (2k+1)\theta_{k+1} \leq k^2(\theta_{k+1} - \theta_k) + 2(k+1)\theta_{k+1},$$

and so

$$\Big[k^2 d_k + (k+1)^2\theta_{k+1}\Big] - \Big[(k-1)^2 d_{k-1} + k^2\theta_k\Big] \leq 2(k+1)\theta_{k+1}. \tag{16}$$

The result is obtained by observing that $k^2 d_k + (k+1)^2\theta_{k+1}$ is bounded from below and the right-hand side of (16) is summable (by Fact 3). □

We are now in a position to prove Theorem 1.

Proof of Theorem 1. From Facts 3 and 4, we deduce that

$$\sum_{k=1}^{\infty}\frac{1}{k}\Big[k^2\|x_{k+1} - x_k\|^2 + (k+1)^2\big(\Theta(x_{k+1}) - \Theta(x^*)\big)\Big] < +\infty.$$

Combining this with Lemma 2, we obtain

$$\lim_{k\to\infty}\Big[k^2\|x_{k+1} - x_k\|^2 + (k+1)^2\big(\Theta(x_{k+1}) - \Theta(x^*)\big)\Big] = 0.$$

Since all the terms are nonnegative, we conclude that both limits are 0, as claimed. ■

Remark 2. Facts 3 and 4, also imply that the function values and the velocities satisfy

$$\liminf_{k\to\infty} k^2\ln(k)\Big((\Psi+\Phi)(x_k)-\min(\Psi+\Phi)\Big)=0 \qquad \text{and} \qquad \liminf_{k\to\infty} k\ln(k)\|x_{k+1}-x_k\|=0,$$

respectively. Indeed, if β_k is any nonnegative sequence such that $\sum_{k=1}^{\infty}\frac{\beta_k}{k}<\infty$ (which holds for (k^2d_k) and $(k^2\theta_k)$), then it cannot be true that $\liminf_{k\to\infty}\beta_k\ln(k)\geq\varepsilon>0$. Otherwise, $\frac{\beta_k}{k}\geq\frac{\varepsilon}{k\ln(k)}$ for all sufficiently large k, and the series above would be divergent.

1.3. **Convergence of the sequence.** It is possible to prove that the sequences generated by (2) converge weakly to minimizers of $\Psi+\Phi$ when $\alpha>3$. Although this was already shown in [2], we provide a proof following the preceding ideas, for completeness.

Theorem 3. *Let $\Psi:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ be proper, lower-semicontinuous and convex, and let $\Phi:\mathcal{H}\to\mathbb{R}$ be convex and continuously differentiable with L-Lipschitz continuous gradient. Suppose that $S=\operatorname{argmin}(\Psi+\Phi)\neq\emptyset$, and let (x_k) be a sequence generated by algorithm* (2) *with $\alpha>3$ and $0<s<\frac{1}{L}$. Then, the sequence (x_k) converges weakly to a point in S.*

Proof. Using the definition (6) of z_k, we write

$$\begin{aligned}\|z_k-x^*\|^2 &= \left(\frac{k-1}{\alpha-1}\right)^2\|x_k-x_{k-1}\|^2+2\frac{k-1}{\alpha-1}\langle x_k-x^*,x_k-x_{k-1}\rangle+\|x_k-x^*\|^2\\ &= \left[\left(\frac{k-1}{\alpha-1}\right)^2+\left(\frac{k-1}{\alpha-1}\right)\right]\|x_k-x_{k-1}\|^2+\left(\frac{k-1}{\alpha-1}\right)\Big[\|x_k-x^*\|^2-\|x_{k-1}-x^*\|^2\Big]+\|x_k-x^*\|^2.\end{aligned}$$

We shall prove that $\lim_{k\to\infty}\|z_k-x^*\|$ exists. By Lemma 2 (or Theorem 1) and Fact 4, it suffices to prove that

$$\delta_k:=(k-1)\Big[\|x_k-x^*\|^2-\|x_{k-1}-x^*\|^2\Big]+(\alpha-1)\|x_k-x^*\|^2$$

has a limit as $k\to\infty$. Clearly, (δ_k) is bounded, by Facts 2 and 4. Write $h_k:=\|x_k-x^*\|^2$ and notice that

$$\begin{aligned}\delta_{k+1}-\delta_k &= (\alpha-1)(h_{k+1}-h_k)+k(h_{k+1}-h_k)-(k-1)(h_k-h_{k-1})\\ &= (k+\alpha-1)(h_{k+1}-h_k)-(k-1)(h_k-h_{k-1}).\end{aligned} \tag{17}$$

On the other hand, from (11), we obtain

$$\Theta(x_{k+1})-\Theta(x^*)\leq\langle G_s(y_k),y_k-x^*\rangle-\frac{s}{2}\|G_s(y_k)\|^2.$$

Since $x_{k+1}=y_k-sG_s(y_k)$, we have

$$\begin{aligned}0 &\leq 2\langle y_k-x_{k+1},y_k-x^*\rangle-\|y_k-x_{k+1}\|^2\\ &= \|y_k-x_{k+1}\|^2+\|y_k-x^*\|^2-\|x_{k+1}-x^*\|^2-\|y_k-x_{k+1}\|^2,\end{aligned}$$

and so

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq \|y_k-x^*\|^2\\ &= \left\|x_k-x^*+\frac{k-1}{k+\alpha-1}(x_k-x_{k-1})\right\|^2\\ &= \|x_k-x^*\|^2+\left(\frac{k-1}{k+\alpha-1}\right)^2\|x_k-x_{k-1}\|^2+2\frac{k-1}{k+\alpha-1}\langle x_k-x^*,x_k-x_{k-1}\rangle\\ &= \|x_k-x^*\|^2+\left[\left(\frac{k-1}{k+\alpha-1}\right)^2+\frac{k-1}{k+\alpha-1}\right]\|x_k-x_{k-1}\|^2+\frac{k-1}{k+\alpha-1}\Big[\|x_{k+1}-x^*\|^2-\|x_k-x^*\|^2\Big]\\ &\leq \|x_k-x^*\|^2+2\|x_k-x_{k-1}\|^2+\frac{k-1}{k+\alpha-1}\Big[\|x_k-x^*\|^2-\|x_{k-1}-x^*\|^2\Big].\end{aligned}$$

In other words,

$$(k+\alpha-1)(h_{k+1}-h_k)-(k-1)(h_k-h_{k-1})\leq 2(k+\alpha-1)\|x_k-x_{k-1}\|^2.$$

Injecting this in (17), we deduce that

$$\delta_{k+1}-\delta_k\leq 2(k+\alpha-1)\|x_k-x_{k-1}\|^2.$$

Since the right-hand side is summable and (δ_k) is bounded, $\lim_{k\to\infty}\delta_k$ exists. It follows that $\lim_{k\to\infty}\|z_k-x^*\|$ exists. In view of Theorem 1 and the definition (6) of z_k, $\lim_{k\to\infty}\|x_k-x^*\|$ exists. Since this holds for any $x^*\in S$, Opial's Lemma shows that the sequence (x_k) converges weakly, as $k\to+\infty$, to a point in S. $\square$

1.4. **Stability under additive errors.** Consider the inexact version of Algorithm (2) given by

$$\left\{\begin{array}{rcl} y_k &=& x_k + \frac{k-1}{k+\alpha-1}(x_k - x_{k-1}) \\ \\ x_{k+1} &=& \operatorname{prox}_{s\Phi}\left(y_k - s(\nabla\Psi(y_k) - g_k)\right). \end{array}\right. \tag{18}$$

The second relation means that

$$y_k - s\nabla\Psi(y_k) \in x_{k+1} + s\Big(\partial\Phi(x_{k+1}) + B(0,\varepsilon_{k+1})\Big)$$

for any $\varepsilon_{k+1} > \|g_k\|$. It turns out that it is possible to give a tolerance estimation for the sequence of errors (g_k) in order to ensure that all the asymptotic properties of (2) (including the $o(k^{-2})$ order of convergence) hold for (18). More precisely, we have the following:

Theorem 4. *Let $\Psi : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ be proper, lower-semicontinuous and convex, and let $\Phi : \mathcal{H} \to \mathbb{R}$ be convex and continuously differentiable with L-Lipschitz continuous gradient. Suppose that $S = \operatorname{argmin}(\Psi+\Phi) \neq \emptyset$, and let (x_k) be a sequence generated by algorithm (18) with $\alpha > 3$ and $0 < s < \frac{1}{L}$. If $\sum_{k=1}^{\infty} k\|g_k\| < +\infty$, then, the function values and the velocities satisfy $\lim_{k\to\infty} k^2\Big((\Psi+\Phi)(x_k) - \min(\Psi+\Phi)\Big) = 0$ and $\lim_{k\to\infty} k\|x_{k+1}-x_k\| = 0$, respectively. Moreover, (x_k) converges weakly to a point in S.*

The key idea is to observe that, for each $k \geq 1$, we have

$$\mathcal{E}(k) \leq \mathcal{E}(0) + \sum_{j=0}^{k-1} 2s\,(j+\alpha-1)\,\langle g_j, z_{j+1} - x^*\rangle$$

(with the same definitions of z_k and $\mathcal{E}(k)$ given in (6) and (5), respectively). This implies

$$\|z_k - x^*\|^2 \leq \frac{1}{\alpha-1}\mathcal{E}(0) + \frac{2s}{\alpha-1}\sum_{j=1}^{k}(j+\alpha-2)\,\|g_{j-1}\|\|z_j - x^*\|.$$

Then, we apply Lemma [2, Lemma A.9] with $a_k = \|z_k - x^*\|$ to deduce that the sequence (z_k) is bounded and so, the modified energy sequence $(\mathcal{F}(k))$, given by

$$\mathcal{F}(k) := \frac{2s}{\alpha-1}\,(k+\alpha-2)^2\,(\Theta(x_k) - \Theta(x^*) + (\alpha-1)\|z_k - x^*\|^2 + \sum_{j=k}^{\infty} 2s\,(j+\alpha-1)\,\langle g_j, z_{j+1} - x^*\rangle\,,$$

is well defined and nonincreasing. The rest of the proof follows pretty much the arguments given above with $\mathcal{E}$ replaced by $\mathcal{F}$ (see also [2, Section 5]).

Inexact FISTA-like algorithms have also been considered in [23, 24]. It would be interesting to obtain similar order-of-convergence results under *relative error* conditions.

Acknowledgement. The authors thank Patrick Redont for his valuable remarks.

References

[1] F. Alvarez, H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Analysis, 9 (2001), No. 1-2, pp. 3–11.

[2] H. Attouch, Z. Chbani, J. Peypouquet, P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing damping*, Paper under review.

[3] H. Bauschke, P. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics, Springer, (2011).

[4] A. Beck, M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), No. 1, pp. 183–202.

[5] A. Chambolle, C. Dossal, *On the convergence of the iterates of Fista*, HAL Id: hal-01060130 https://hal.inria.fr/hal-01060130v3 Submitted on 20 Oct 2014.

[6] A. Chambolle, T. Pock, *A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions*, SMAI Journal of Computational Mathematics 1 (2015), pp. 29–54.

[7] P.L. Combettes, V.R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[8] I. Daubechies, M. Defrise, C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.

[9] A.A. Goldstein, *Convex programming in Hilbert space*, Bulletin of the American Mathematical Society 70 (1964) pp. 709–710.

[10] E.S. Levitin, B.T. Polyak, *Constrained minimization problems*, USSR Computational Mathematics and Mathematical Physics 6 (1966) pp. 1–50.

[11] P.L. Lions, B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[12] R. May, *Asymptotic for a second order evolution equation with convex potential and vanishing damping term*, arXiv:1509.05598.

[13] A. Moudafi, M. Oliny, *Convergence of a splitting inertial proximal method for monotone operators*, J. Comput. Appl. Math., 155 (2003), No. 2, pp. 447–454.

[14] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

[15] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[16] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical programming, 103 (2005), No. 1, pp. 127–152.

[17] Y. Nesterov, *Gradient methods for minimizing composite objective function*, CORE Discussion Papers, 2007.

[18] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[19] N. Parikh, S. Boyd, *Proximal algorithms*, Foundations and trends in optimization, volume 1, (2013), pp. 123-231.

[20] G.B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390.

[21] J. Peypouquet, *Convex optimization in normed spaces: theory, methods and examples.* Springer, 2015.

[22] D.A. Lorenz, T. Pock, *An inertial forward-backward algorithm for monotone inclusions*, J. Math. Imaging Vision, pp. 1-15, 2014. (online).

[23] M. Schmidt, N. Le Roux, F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, NIPS'11 - 25 th Annual Conference on Neural Information Processing Systems, Dec 2011, Grenada, Spain. (2011) HAL inria-00618152v3.

[24] S. Villa, S. Salzo, L. Baldassarres, A. Verri, *Accelerated and inexact forward-backward*, SIAM J. Optim., 23 (2013), No. 3, pp. 1607–1633.

[25] W. Su, S. Boyd, E.J. Candès, *A Differential equation for modeling Nesterov's accelerated gradient method: theory and insights.* Neural Information Processing Systems (NIPS) 2014.

Institut de Mathématiques et Modélisation de Montpellier, UMR 5149 CNRS, Université Montpellier 2, place Eugène Bataillon, 34095 Montpellier cedex 5, France
E-mail address: hedy.attouch@univ-montp2.fr

Departamento de Matemática, Universidad Técnica Federico Santa María, Av España 1680, Valparaíso, Chile
E-mail address: juan.peypouquet@usm.cl