

THE RATE OF CONVERGENCE OF NESTEROV'S ACCELERATED FORWARD-BACKWARD METHOD IS ACTUALLY FASTER THAN $1/k^{2*}$

HEDY ATTOUCH[†] AND JUAN PEYPOUQUET[‡]

Abstract. The *forward-backward algorithm* is a powerful tool for solving optimization problems with an *additively separable* and *smooth* plus *nonsmooth* structure. In the convex setting, a simple but ingenious acceleration scheme developed by Nesterov improves 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.

Key words. convex optimization, fast convergent methods, Nesterov method

AMS subject classifications. 49M37, 65K05, 90C25

DOI. 10.1137/15M1046095

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

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

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

The *forward-backward* method, which generalizes the *gradient projection* algorithm [9, 11], was proposed in [12] and [21] to overcome the inherent difficulties of minimizing the nonsmooth sum of two functions, as in (1), while exploiting its *additively separable* and *smooth* plus *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-thresholding algorithm* (ISTA). Some time later, a decisive improvement came with [4], where ISTA was successfully combined with Nesterov's acceleration scheme [15] to produce the *fast iterative shrinkage-thresholding algorithm* (FISTA). For general Φ and Ψ , and after some simplification, the *accelerated forward-backward* method can be written as the following iteration:

$$(2) \quad \begin{cases} y_k = x_k + \frac{k-1}{k+\alpha-1}(x_k - x_{k-1}), \\ x_{k+1} = \text{prox}_{s\Psi}(y_k - s(\nabla\Phi(y_k))), \end{cases}$$

*Received by the editors October 29, 2015; accepted for publication (in revised form) July 11, 2016; published electronically September 1, 2016. This effort was sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant FA9550-14-1-0056, and was 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. <http://www.siam.org/journals/siopt/26-3/M104609.html>

[†]Institut de Mathématiques et Modélisation de Montpellier, UMR 5149 CNRS, Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France (hedy.attouch@univ-montp2.fr).

[‡]Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile (juan.peypouquet@usm.cl, <http://jpeypou.mat.utfsm.cl/>).

where $\alpha > 0$ and $s > 0$, and the points x_0 and x_1 are arbitrarily taken in \mathcal{H} . Here, prox_f denotes the *proximal mapping* of a function f (see [3, Definition 12.23] or [20, Definition 1.1]). This algorithm is closely connected with proximal-based inertial algorithms [1, 14, 23]. 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 convergence rate of

$$(\Psi + \Phi)(x_k) - \min(\Psi + \Phi) = \mathcal{O}(k^{-2}).$$

More precisely, let L be the Lipschitz constant of $\nabla\Phi$ and let D denote the distance from the initial point x_0 to the set of solutions for (1). Setting $s = 1/L$, we have

$$(3) \quad (\Psi + \Phi)(x_k) - \min(\Psi + \Phi) \leq \frac{2LD^2}{(k+1)^2}$$

(see [16], or also [10]). Observe that this bound is uniform with respect to the objective functions. In turn, the convergence rate obtained for the unaccelerated counterpart is just $\mathcal{O}(k^{-1})$.

Nevertheless, while sequences generated by the classical forward-backward method are (weakly) 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, using different arguments. In [5], the authors use a *descent* inequality satisfied by forward-backward iterations (see also [6, section 2.2]). In turn, the proof given in [2] relies on the connection between (2) and the differential inclusion

$$(4) \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial\Psi(x(t)) + \nabla\Phi(x(t)) \ni 0.$$

Indeed, as pointed out in [26, 2], algorithm (2) can be seen as an appropriate finite-difference discretization of (4). In [26], the authors studied the equation

$$(5) \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\Theta(x(t)) = 0,$$

where Θ is a smooth convex function on \mathcal{H} , and proved that

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

when $\alpha \geq 3$. Most of our arguments (as well as those in [2]) are inspired by their analysis. 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 based on this relationship. If $\alpha > 3$, we actually have

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

which means that

$$\lim_{t \rightarrow \infty} t^2(\Theta(x(t)) - \min \Theta) = 0.$$

Although it can be derived from the arguments in [2], it was May [13] 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 $\mathcal{O}(k^{-2})$, namely $o(k^{-2})$.

At several points in this paper, we shall make the following set of assumptions.

HYPOTHESIS (H). The function $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper, lower semicontinuous, and convex, and the function $\Phi : \mathcal{H} \rightarrow \mathbb{R}$ is convex and continuously differentiable with L -Lipschitz continuous gradient. The set $S = \operatorname{argmin}(\Psi + \Phi)$ is nonempty and $0 < s \leq \frac{1}{L}$.

The main result of this paper is the following.

THEOREM 1. *Let Hypothesis (H) hold and let (x_k) be a sequence generated by algorithm (2) with $\alpha > 3$. Then,*

$$(6) \quad \lim_{k \rightarrow \infty} k^2 \left((\Psi + \Phi)(x_k) - \min(\Psi + \Phi) \right) = 0 \quad \text{and} \quad \lim_{k \rightarrow \infty} k \|x_{k+1} - x_k\| = 0.$$

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

Moreover, we recover some results from [2, section 5], closely connected with the ones in [5], with simplified arguments.

Some comments are in order: First, 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. Second, Nesterov [15] gave an example of a function $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}$ for which the algorithm described by (2) (with $\Psi \equiv 0$) would satisfy

$$(7) \quad \Phi(x_k) - \min \Phi \geq \frac{3D^2}{32(k+1)^2},$$

as long as $k \leq (N-1)/2$, where, as in (3), D denotes the distance from the initial point x_0 to the set of solutions for (1) (see [16, Theorem 2.1.7] or [10, section 3.3]). Theorem 1 does not contradict this fact, and implies that inequality (7) cannot hold for all k .

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 [24, 25].

1. Main results. Throughout this section, Hypothesis (H) is in force, and the sequence (x_k) is generated by algorithm (2) with $\alpha \geq 3$. To simplify the notation, we set $\Theta = \Psi + \Phi$. For standard convex analysis background, see [3, 22].

1.1. Some important estimations. We begin by establishing the basic properties of the sequence (x_k) . Some results can be found in [2, 5] (especially, Facts 2 and 3), for which we provide simplified proofs. As mentioned above, many arguments can be traced back to [26] (see, for instance, the definition of \mathcal{E} given in (8) and (9)).

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

$$(8) \quad \mathcal{E}(k) := \frac{2s}{\alpha-1} (k+\alpha-2)^2 (\Theta(x_k) - \Theta(x^*)) + (\alpha-1) \|z_k - x^*\|^2,$$

where

$$(9) \quad z_k := \frac{k+\alpha-1}{\alpha-1} y_k - \frac{k}{\alpha-1} x_k = x_k + \frac{k-1}{\alpha-1} (x_k - x_{k-1}).$$

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

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

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

$$(10) \quad x_{k+1} = y_k - sG_s(y_k).$$

The variable z_k , defined in (9), will play an important role. Simple algebraic manipulations give

$$(11) \quad z_{k+1} = \frac{k+\alpha-1}{\alpha-1} (y_k - sG_s(y_k)) - \frac{k}{\alpha-1} x_k = z_k - \frac{s}{\alpha-1} (k+\alpha-1) G_s(y_k).$$

The operator G_s satisfies

$$(12) \quad \Theta(y - sG_s(y)) \leq \Theta(x) + \langle G_s(y), y - x \rangle - \frac{s}{2} \|G_s(y)\|^2$$

for all $x, y \in \mathcal{H}$ (see [4, 5, 20, 26]), 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

$$(13) \quad \Theta(y_k - sG_s(y_k)) \leq \Theta(x_k) + \langle G_s(y_k), y_k - x_k \rangle - \frac{s}{2} \|G_s(y_k)\|^2$$

and

$$(14) \quad \Theta(y_k - sG_s(y_k)) \leq \Theta(x^*) + \langle G_s(y_k), y_k - x^* \rangle - \frac{s}{2} \|G_s(y_k)\|^2,$$

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

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

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

$$(15) \quad \begin{aligned} \Theta(x_{k+1}) &\leq \frac{k}{k+\alpha-1} \Theta(x_k) + \frac{\alpha-1}{k+\alpha-1} \Theta(x^*) \\ &\quad + \frac{\alpha-1}{k+\alpha-1} \langle G_s(y_k), z_k - x^* \rangle - \frac{s}{2} \|G_s(y_k)\|^2. \end{aligned}$$

We shall obtain a recursion from (15). To this end, observe that (11) gives

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

After developing

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

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

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

Replacing this in (15), we deduce that

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

Equivalently,

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

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

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

which implies

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

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,

$$(16) \quad \mathcal{E}(k+1) + 2s\frac{\alpha-3}{\alpha-1}k (\Theta(x_k) - \Theta(x^*)) \leq \mathcal{E}(k).$$

We then deduce the following facts.

Fact 1. The sequence $(\mathcal{E}(k))$ is nonincreasing and $\lim_{k \rightarrow \infty} \mathcal{E}(k)$ 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 (16), we also obtain the following fact.

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

Now, using (13) 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

$$(17) \quad \Theta(x_{k+1}) + \frac{1}{2s} \|x_{k+1} - x_k\|^2 \leq \Theta(x_k) + \frac{1}{2s} \frac{(k-1)^2}{(k+\alpha-1)^2} \|x_k - x_{k-1}\|^2.$$

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 (17) as

$$(18) \quad \theta_{k+1} + d_k \leq \theta_k + \frac{(k-1)^2}{(k+\alpha-1)^2} d_{k-1}.$$

Since $k + \alpha - 1 \geq k + 1$, (18) 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 $k = 1, \dots, K$, 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. Using this inequality, along with (16) (with $k = 1$), we obtain the following result.

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

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

1.2. From $\mathcal{O}(k^{-2})$ to $o(k^{-2})$. Recall that $\Psi : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper, lower semicontinuous, and convex, $\Phi : \mathcal{H} \rightarrow \mathbb{R}$ 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 $0 < s \leq \frac{1}{L}$. We shall prove that (6) holds; in other words, that $(\Psi + \Phi)(x_k) - \min(\Psi + \Phi) = o(k^{-2})$ and $\|x_{k+1} - x_k\| = o(k^{-1})$.

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

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

Proof. Since $k + \alpha - 1 \geq k$, inequality (18) 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

$$(19) \quad \left[k^2 d_k + (k+1)^2 \theta_{k+1} \right] - \left[(k-1)^2 d_{k-1} + k^2 \theta_k \right] \leq 2(k+1) \theta_{k+1}.$$

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 (19) is summable (by Fact 3). \square

We are now in a position to prove Theorem 1.

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

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

Combining this with Lemma 2, we obtain

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

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

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

$$\liminf_{k \rightarrow \infty} k^2 \ln(k) \left((\Psi + \Phi)(x_k) - \min(\Psi + \Phi) \right) = 0 \quad \text{and} \quad \liminf_{k \rightarrow \infty} k \ln(k) \|x_{k+1} - x_k\| = 0,$$

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

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, 5], we provide a proof following the preceding ideas, for completeness.

THEOREM 3. *Let Hypothesis (H) hold, and let (x_k) be a sequence generated by algorithm (2) with $\alpha > 3$. Then, the sequence (x_k) converges weakly to a point in S .*

Proof. Take any $x^* \in S$. Using the definition (9) of z_k , we write

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

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

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

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

$$\begin{aligned}\delta_{k+1} - \delta_k &= (\alpha-1)(h_{k+1} - h_k) + k(h_{k+1} - h_k) - (k-1)(h_k - h_{k-1}) \\ (20) \quad &= (k+\alpha-1)(h_{k+1} - h_k) - (k-1)(h_k - h_{k-1}).\end{aligned}$$

On the other hand, from (14), 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

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

In other words,

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

Substituting this in (20), 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, $\lim_{k \rightarrow \infty} \delta_k$ exists. It follows that $\lim_{k \rightarrow \infty} \|z_k - x^*\|$ exists. In view of Theorem 1 and the definition (9) of z_k , $\lim_{k \rightarrow \infty} \|x_k - x^*\|$ exists. Since this holds for any $x^* \in S$, Opial's Lemma [19] (see, for instance, [22, Lemma 5.2]) shows that the sequence (x_k) converges weakly, as $k \rightarrow +\infty$, to a point in S . \square

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

$$(21) \quad \begin{cases} y_k = x_k + \frac{k-1}{k+\alpha-1}(x_k - x_{k-1}), \\ x_{k+1} = \text{prox}_{s\Phi}(y_k - s(\nabla\Psi(y_k) - g_k)). \end{cases}$$

The second relation means that

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

for any $\varepsilon_{k+1} \geq \|g_k\|$. 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 (21). More precisely, we have the following:

THEOREM 4. *Let Hypothesis **(H)** hold, and let (x_k) be a sequence generated by algorithm (21) with $\alpha > 3$. If $\sum_{k=1}^{\infty} k\|g_k\| < +\infty$, then, $\lim_{k \rightarrow \infty} k^2((\Psi + \Phi)(x_k) - \min(\Psi + \Phi)) = 0$ and $\lim_{k \rightarrow \infty} k\|x_{k+1} - x_k\| = 0$. 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 (9) and (8), 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

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

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

Inexact FISTA-like algorithms have also been considered in [24, 25]. It would be interesting to obtain similar order-of-convergence results under *relative error* conditions.

Acknowledgments. The authors thank Patrick Redont and the anonymous referees for their valuable remarks.

REFERENCES

- [1] F. ALVAREZ AND H. ATTOUCH, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Var. Anal., 9 (2001), No. 1-2, pp. 3–11.
- [2] H. ATTOUCH, Z. CHBANI, J. PEYPOUQUET, AND P. REDONT, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing damping*, Math. Program., to appear, DOI 10.1007/s10107-016-0992-8.
- [3] H. BAUSCHKE AND P. COMBETTES, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics, Springer, New York, 2011.
- [4] A. BECK AND M. TEOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imag. Sci., 2 (2009), No. 1, pp. 183–202.
- [5] A. CHAMBOLLE AND C. DOSSAL, *On the convergence of the iterates of Fista*, preprint, hal-01060130, <https://hal.inria.fr/hal-01060130v3>.
- [6] A. CHAMBOLLE AND T. POCK, *A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions*, SMAI J. Comput. Math., 1 (2015), pp. 29–54.
- [7] P. L. COMBETTES AND V. R. WAJS, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
- [8] I. DAUBECHIES, M. DEFRISE, AND C. DE MOL, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure Appl. Math., 57 (2004), pp. 1413–1457.
- [9] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Am. Math. Soc., 70 (1964), pp. 709–710.
- [10] D. KIM AND J. A. FESSLER, *Optimized first-order methods for smooth convex minimization*, Math. Program., Series A, 159 (2016), pp. 81–107.
- [11] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization problems*, USSR Comput. Math. Math. Phys. 6 (1966), pp. 1–50.
- [12] P. L. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
- [13] R. MAY, *Asymptotic for a second order evolution equation with convex potential and vanishing damping term*, preprint, arXiv:1509.05598.
- [14] A. MOUDAFI AND M. OLINY, *Convergence of a splitting inertial proximal method for monotone operators*, J. Comput. Appl. Math., 155 (2003), No. 2, pp. 447–454.
- [15] Y. NESTEROV, *A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$* , Dokl. Akad. Nauk SSSR, 27 (1983), pp. 372–376.
- [16] Y. NESTEROV, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization, Kluwer Academic, Boston, MA, 2004.
- [17] Y. NESTEROV, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), No. 1, pp. 127–152.
- [18] Y. NESTEROV, *Gradient methods for minimizing composite objective function*, CORE Discussion Papers, No. 76, Universite Catholique de Louvain, 2007.
- [19] Z. OPIAL, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Am. Math. Soc., 73 (1967), pp. 591–597.
- [20] N. PARIKH AND S. BOYD, *Proximal algorithms*, Found. Trends Optim. 1 (2013), pp. 123–231.
- [21] 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.
- [22] J. PEYPOUQUET, *Convex optimization in normed spaces: Theory, methods and examples*. Springer, New York 2015.

- [23] D. A. LORENZ AND T. POCK, *An inertial forward-backward algorithm for monotone inclusions*, J. Math. Imag. Vis., (2014), pp. 1–15; also available (online).
- [24] M. SCHMIDT, N. LE ROUX, AND F. BACH, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Advances in Neural Information Processing System 24, NIPS 2011.
- [25] S. VILLA, S. SALZO, L. BALDASSARRES, AND A. VERRI, *Accelerated and inexact forward-backward*, SIAM J. Optim., 23 (2013), No. 3, pp. 1607–1633.
- [26] W. SU, S. BOYD, AND E. J. CANDÈS, *A Differential equation for modeling Nesterov's accelerated gradient method: Theory and insights* in Advances in Neural Information Processing System 27, NIPS 2014.