SIAM J. OPTIM.
Vol. 30, No. 3, pp. 2134–2162
© 2020 Society for Industrial and Applied Mathematics

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

FINITE CONVERGENCE OF PROXIMAL-GRADIENT INERTIAL ALGORITHMS COMBINING DRY FRICTION WITH HESSIAN-DRIVEN DAMPING*

SAMIR ADLY† AND HEDY ATTOUCH‡

Abstract. In a Hilbert space $\mathcal{H}$, we introduce a new class of proximal-gradient algorithms with finite convergence properties. These algorithms naturally occur as discrete temporal versions of an inertial differential inclusion which is stabilized under the joint action of three dampings: dry friction, viscous friction, and a geometric damping driven by the Hessian. The function $f : \mathcal{H} \to \mathbb{R}$ to be minimized is supposed to be differentiable (not necessarily convex) and enters the algorithm via its gradient. The dry friction damping function $\phi : \mathcal{H} \to \mathbb{R}_+$ is convex with a sharp minimum at the origin (typically $\phi(x) = r\|x\|$ with $r > 0$). It enters the algorithm via its proximal mapping, which acts as a soft threshold operator on the velocities. It is the source of stabilization in a finite number of steps. The geometric damping driven by the Hessian intervenes in the dynamics in the form $\nabla^2 f(x(t))\dot{x}(t)$. By treating this term as the time derivative of $\nabla f(x(t))$, this gives, in discretized form, first-order algorithms. The Hessian-driven damping allows one to control and to attenuate the oscillations. The convergence results tolerate the presence of errors, under the sole assumption of their asymptotic convergence to zero. Replacing the potential function f by its Moreau envelope, we extend the results to the case of a nonsmooth convex function f. In this case, the algorithm involves the proximal operators of f and ϕ separately. Several variants of this algorithm are considered, including the case of the Nesterov accelerated gradient method. We then consider the extension to the case of additive composite optimization, thus leading to splitting methods. Numerical experiments are given for lasso-type problems. The performance profiles, as a comparison tool, highlight the effectiveness of a variant of the Nesterov accelerated method with dry friction and Hessian-driven damping.

Key words. proximal-gradient algorithms, inertial methods, differential inclusion, dry friction, Hessian-driven damping, finite convergence, lasso problem

AMS subject classifications. 37N40, 34A60, 34G25, 49K24, 70F40

DOI. 10.1137/19M1307779

1. Introduction and preliminary results.

Throughout the paper $\mathcal{H}$ is a real Hilbert space, with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$, and $f : \mathcal{H} \to \mathbb{R}$ is a $\mathcal{C}^1$ function whose gradient is Lipschitz continuous. When we consider continuous dynamics on which the algorithms are based, and where the Hessian intervenes, more regularity is needed for f which is then assumed to be a $\mathcal{C}^2$ function. Several extensions of these hypotheses will be discussed later in the paper.

1.1. Presentation of the algorithm and main contributions.

We analyze the finite convergence (within a finite number of steps) of several algorithms that can be obtained by different temporal discretizations of the second-order differential inclusion

$$\text{(IGDH)} \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \partial\phi(\dot{x}(t)) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0, \; t \in [t_0, +\infty[,$$

where γ and β are positive damping parameters. (IGDH) stands for inertial gradient system with dry friction and Hessian-driven damping. The dry friction damping

*Received by the editors December 18, 2019; accepted for publication (in revised form) June 16, 2020; published electronically August 11, 2020.
https://doi.org/10.1137/19M1307779

†Laboratoire XLIM, Université de Limoges, 87060 Limoges Cedex, France (samir.adly@unilim.fr).
‡IMAG, Université de Montpellier, CNRS, 34095 Montpellier Cedex 5, France (hedy.attouch@univ-montp2.fr).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

function $\phi : \mathcal{H} \to \mathbb{R}_+$ is convex with a sharp minimum at the origin, typically $\phi(x) = r\|x\|$ with $r > 0$. The geometric damping driven by the Hessian intervenes in the dynamics in the form $\nabla^2 f(x(t))\dot{x}(t)$. By treating this term as the time derivative of $\nabla f(x(t))$, this gives, in discretized form, first-order algorithms.

Our goal in this paper is to focus on various temporal discretizations of (IGDH) and their links with numerical optimization. For the well-posedness and the finite-time stabilization of the continuous dynamic (IGDH), we refer to [3].

Our main results concern the finite convergence of the inertial proximal-gradient algorithm with Hessian-damping and dry friction

$$\text{(IPAHDD)} \quad \begin{cases} z_k = \dfrac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \dfrac{\beta}{1+h\gamma}(\nabla f(x_k) - \nabla f(x_{k-1})) - \dfrac{h}{1+h\gamma}\nabla f(x_k) \\ x_{k+1} = x_k + h\,\mathrm{prox}_{\frac{h}{1+h\gamma}\phi}(z_k), \end{cases}$$

which comes from the temporal discretization with step size $h > 0$ of (IGDH). In the above formula, $\mathrm{prox}_\phi : \mathcal{H} \to \mathcal{H}$ denotes the proximal mapping associated with the convex function ϕ. Recall that, for any $x \in \mathcal{H}$, for any $\lambda > 0$

$$\mathrm{prox}_{\lambda\phi}(x) := \mathrm{argmin}_{\xi\in\mathcal{H}} \left\{ \lambda\phi(\xi) + \frac{1}{2}\|x - \xi\|^2 \right\}.$$

Assuming that the viscous damping parameter γ is taken large enough, we show in Theorem 2.1 that any sequence (x_k) generated by the algorithm (IPAHDD) satisfies the summability property

$$\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\| < +\infty.$$

This property expresses that the trajectory has a finite length, and therefore $\lim_{k\to\infty} x_k := x_\infty$ exists for the strong topology of $\mathcal{H}$. The limit point x_∞ satisfies

$$-\nabla f(x_\infty) \in \partial\phi(0). \tag{1.1}$$

It is an "approximate" critical point of f. Equation (1.1) is the stationary property for the evolution equation (IGDH) (make $\dot{x}$ and $\ddot{x}$ equal to zero). Since our goal is to minimize the function f, we have to choose a function ϕ whose subdifferential set $\partial\phi(0)$ is "relatively small." Let us illustrate this in the model situation $\phi(x) = r\|x\|$. Property (1.1) is then written[1]

$$\|\nabla f(x_\infty)\| \le r,$$

which leads to taking r a small positive number. Recall that f is not supposed to be convex, in which case the above condition is a standard stopping criterion for optimization algorithms. In addition, we show (Theorem 2.2) that, under the following condition (which is a reinforced version of (1.1))

$$-\nabla f(x_\infty) \in \mathrm{int}(\partial\phi(0)),$$

there is finite convergence (i.e., within a finite number of steps) of the iterates generated by the algorithm (IPAHDD). In short, dry friction acts as a closed-loop stopping

[1]This amounts to solving the optimization problem $\min_{\mathcal{H}} f$ with the Ekeland variational principle, instead of the Fermat rule.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

rule. The Hessian-driven damping allows us to control and to attenuate the oscillation effects which occur naturally with the inertial systems, and which are not desirable from the optimization point of view.

We show that various discretizations of the dynamic (IGDH) lead to different algorithms which share similar convergence properties, including the combination of dry friction and Hessian-driven damping with the accelerated gradient method of Nesterov (we refer to Theorem 3.1 for more details). To the best of our knowledge, this is the first time that the combination of the three types of friction (viscous, Hessian damping, and dry friction) has been used to design efficient optimization algorithms, which makes this contribution new and original in the literature.

1.2. Some historical facts. Let us explain the role and the importance of each of the three damping terms which enter the continuous dynamics (IGDH) and which consequently play a central role in the optimization properties of the associated algorithm (IPAHDD).

1.2.1. Viscous friction. Polyak initiated the use of inertial dynamics to accelerate the gradient method in optimization. In [41], based on the inertial system with a fixed viscous damping coefficient $\gamma > 0$,

$$\text{(HBF)} \qquad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0,$$

he introduced the heavy ball with friction method. For a strongly convex function f, and γ judiciously chosen, (HBF) provides convergence at an exponential rate of $f(x(t))$ to $\min_{\mathcal{H}} f$. For general convex functions, the asymptotic convergence rate of (HBF) is $\mathcal{O}(\frac{1}{t})$ (in the worst case). For a recent account on the convergence results in the case of a time-dependent viscous damping coefficient $\gamma(t)$, we refer to [10], where (HBF) is considered precisely in [10, Corollary 4.5]. A decisive step to improve (HBF) was taken by Su, Boyd, and Candès [45] with the introduction of an asymptotic vanishing damping coefficient $\gamma(t) = \frac{\alpha}{t}$, that is,

$$\text{(AVD)}_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0.$$

As a specific feature, the viscous damping coefficient $\frac{\alpha}{t}$ vanishes (tends to zero) as time t goes to infinity, hence the terminology. For general convex functions, it provides a continuous version of the accelerated gradient method of Nesterov [39, 40]. For $\alpha \geq 3$, each trajectory $x(\cdot)$ of (AVD)_α satisfies the asymptotic convergence rate of the values $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(1/t^2\right)$. The convergence properties of the dynamic (AVD)_α have been the subject of many recent studies; see [8, 10, 11, 12, 14, 16, 18, 20, 21, 37, 45]. The case $\alpha = 3$, which corresponds to Nesterov's historical algorithm, is critical. In the case $\alpha = 3$, the question of the convergence of the trajectories remains an open problem (except in one dimension where convergence holds [16]). For $\alpha > 3$, it has been shown by Attouch et al. [14] that each trajectory converges weakly to a minimizer. The corresponding algorithmic result has been obtained by Chambolle and Dossal [30]. For $\alpha > 3$, it is shown in [18] and [37] that the asymptotic convergence rate of the values is actually $o(1/t^2)$. The subcritical case $\alpha \leq 3$ has been examined by Apidopoulos, Aujol, and Dossal [8] and Attouch, Chbani, and Riahi [16], with the convergence rate of the objective values $\mathcal{O}(t^{-\frac{2\alpha}{3}})$. These rates are optimal, that is, they can be reached, or approached arbitrarily close.

1.2.2. Dry friction. The first results concerning the finite convergence property under the action of dry friction have been obtained by Adly, Attouch, and Cabot [4] for the continuous dynamics

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\ddot{x}(t) + \partial\phi(\dot{x}(t)) + \nabla f(x(t)) \ni 0,\ t \in [t_0, +\infty[. \tag{1.2}$$

Assuming that the potential friction function ϕ has a sharp minimum at the origin (dry friction), they showed that, generically with respect to the initial data, the solution trajectories converge in finite-time to equilibria. Amann and Diaz in [7] highlighted the crucial role played by the dry friction property to obtain stabilization in finite-time of the corresponding continuous dynamic. For a linear oscillator which is damped by a potential friction function $\phi(\xi) = \|\xi\|^p$ with $p > 1$, they showed that certain trajectories do not converge in finite-time. Similar results for the corresponding proximal-based algorithms have been obtained by Baji and Cabot [22] and Adly and Attouch [2]. Let's make precise the tools that will be useful for the mathematical analysis of the set-valued term $\partial\phi(\dot{x}(t))$ in (1.2) which models dry friction. The friction potential function ϕ is supposed to satisfy the dry friction property (denoted by (DF))

$$\text{(DF)} \qquad \begin{cases} \phi : \mathcal{H} \to \mathbb{R} \text{ is convex continuous;} \\ \min\limits_{\xi\in\mathcal{H}} \phi(\xi) = \phi(0) = 0; \\ 0 \in \operatorname{int}(\partial\phi(0)). \end{cases}$$

The particular case $\phi = r\|\cdot\|$, with $r > 0$, models dry friction (also called Coulomb friction) in mechanics. The key assumption $0 \in \operatorname{int}(\partial\phi(0))$ expresses that ϕ has a sharp minimum at the origin. This is specified in the following elementary lemma (see [1, Lemma 4.1, p. 83]), where, in item (iv), ϕ^* is the standard Fenchel conjugate of ϕ.

Lemma 1.1. *Let $\phi : \mathcal{H} \to \mathbb{R}$ be a convex continuous function such that $\min_{\xi\in\mathcal{H}} \phi(\xi) = \phi(0) = 0$. Then, the following formulations of the dry friction are equivalent:*
(i) $0 \in \operatorname{int}(\partial\phi(0))$;
(ii) *there exists some $r > 0$ such that $B(0, r) \subset \partial\phi(0)$;*
(iii) *there exists some $r > 0$ such that, for all $\xi \in \mathcal{H}$, $\phi(\xi) \geq r\|\xi\|$;*
(iv) *there exists some $r > 0$ such that the following implication is satisfied:*

$$g \in \mathcal{H}, \quad \|g\| \leq r \Longrightarrow \partial\phi^*(g) \ni 0.$$

The positive parameter $r > 0$ plays a crucial role in our analysis. To enlighten its role, we say that the friction potential function ϕ satisfies the property $(\mathrm{DF})_r$ if ϕ satisfies the dry friction property (DF) with $B(0, r) \subset \partial\phi(0)$. The property (iv) above expresses that when the norm of the force g exerted on the system is less than a threshold $r > 0$, then the system stabilizes, i.e., the velocity $v = 0 \in \partial\phi^*(g)$. This contrasts with the viscous damping that can asymptotically produce many small oscillations. The following lemma will play a key role in showing the finite convergence property. It gives the soft thresholding property satisfied by the proximal operator associated with a function ϕ having a sharp minimum at the origin. It is a direct consequence of Lemma 1.1(iv).

Lemma 1.2. *Let $\phi : \mathcal{H} \to \mathbb{R}$ be a convex continuous function which satisfies the property $(\mathrm{DF})_r$, i.e., $\partial\phi(0) \supset B(0, r)$. Then, the following implication holds: for $\lambda > 0$, and $x \in \mathcal{H}$*

$$\|x\| \leq \lambda r \implies \operatorname{prox}_{\lambda\phi}(x) = 0.$$

1.2.3. Hessian-driven damping. The inertial system

$$(\mathrm{DIN})_{\gamma,\beta} \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0$$

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

was introduced in [6]. In line with (HBF), the positive viscous friction coefficient γ is *fixed*. The Hessian-driven damping makes it possible to neutralize the transversal oscillations that might occur with (HBF), as observed in [6] in the case of the Rosenbrook function. The need to take a geometric damping adapted to f had already been observed by Álvarez [5], who considered the inertial system

$$\ddot{x}(t) + \Gamma\dot{x}(t) + \nabla f(x(t)) = 0,$$

where $\Gamma : \mathcal{H} \to \mathcal{H}$ is a linear positive anisotropic operator. But still this damping operator is fixed. For a general convex function, the Hessian-driven damping in $(\mathrm{DIN})_{\gamma,\beta}$ performs a similar operation in a closed-loop adaptive way. The terminology (DIN) stands for dynamical inertial Newton. It refers to the natural link between this dynamic and the continuous Newton method. Recent studies have been devoted to the study of the inertial dynamic

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0,$$

which combines asymptotic vanishing damping with Hessian-driven damping. The corresponding algorithms involve a correcting term in the Nesterov accelerated gradient method which reduces the oscillatory aspects; see Attouch, Peypouquet, and Redont [19], Attouch et al. [13], and Shi et al. [43]. These dynamics have proven to be versatile and combine well with approximation methods; see recent contributions from Bot and Csetnek [28] and Bot, Csetnek, and László [26].

1.3. Contents. In section 2, after the introduction of the algorithm (IPAHDD), we establish our main results, which concern its convergence properties. We then specialize our results to the case of the soft thresholding of velocities. Additional results are examined in sections 2.6 and 2.7, which concern respectively the study of a variant of (IPAHDD) and the effect of the introduction of perturbations and errors in the algorithm (IPAHDD). In section 3, we proceed with a similar analysis in the case of the Nesterov acceleration method. In section 4, based on the variational properties of Moreau's envelope, we extend these results to the case where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semicontinuous and proper function. In section 5, we extend our analysis to the case of additive composite optimization problems. For Lasso-type problems, we obtain splitting methods with finite convergence properties and compare the performance of the different algorithms considered previously. A comparison with ISTA and FISTA [24] is included in the numerical experiments.[2] We complete the paper with some perspectives.

2. Inertial proximal-gradient algorithms with dry friction and Hessian damping.

In this section, we assume that f is a $\mathcal{C}^1$ function whose gradient is L-Lipschitz continuous. Unless otherwise indicated, no convexity assumption is made on the function f. We consider a splitting algorithm with the finite convergence property, in which the function to be minimized f intervenes via its gradient, and the dry friction function ϕ via its proximal mapping. We denote by γ, β, r three positive real parameters which intervene respectively in the damping as

$$\begin{cases} \gamma > 0 \ \text{ is a viscous damping parameter;} \\ \beta > 0 \ \text{ is attached to the Hessian-driven damping;} \\ r > 0 \ \text{ is a dry friction parameter, that is, } \phi(x) \geq r\|x\| \ \text{ and } \phi(0) = 0. \end{cases}$$

[2] We thank the anonymous reviewer for suggesting it to us.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

2.1. Proximal-gradient algorithms with Hessian damping and dry friction. Given a constant time step $h > 0$, we consider the following temporal discretization of (IGDH):

$$\begin{aligned} &\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) \\ &\quad + \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) + \nabla f(x_k) \ni 0. \end{aligned} \tag{2.1}$$

It is implicit with respect to the nonsmooth function ϕ and explicit with respect to the smooth function f. It is in line with the classical proximal-gradient methods that deal with additively structured minimization problems *smooth + nonsmooth*. In our situation, this structure involves the friction terms, hence significant differences! We use the equality $\nabla^2 f(x(t))\dot{x}(t) = \frac{d}{dt}\nabla f(x(t))$, which follows from the classical derivation chain rule. So, the corrective term $\nabla f(x_k) - \nabla f(x_{k-1})$ is directly linked to the temporal discretization of the Hessian-driven damping term. It plays a central role in reducing the oscillatory effects that are attached to inertial systems. Solving (2.1) with respect to x_{k+1} gives the following first-order algorithm where dry friction enters via the potential function ϕ, and the function to be minimized f enters via its gradient:

(IPAHDD)

Initialize: $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$

$$\begin{aligned} x_{k+1} = x_k + h\,\mathrm{prox}_{\frac{h}{1+h\gamma}\phi}\Big(&\tfrac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \tfrac{\beta}{1+h\gamma}(\nabla f(x_k) - \nabla f(x_{k-1})) \\ &- \tfrac{h}{1+h\gamma}\nabla f(x_k)\Big) \end{aligned}$$

We call it (IPAHDD), which stands for inertial proximal-gradient algorithm with Hessian damping and dry friction. Consequently, given x_{k-1} and x_k, (IPAHDD) uniquely determines x_{k+1}. When $\phi = 0$, that is, without dry friction, we obtain the inertial gradient algorithm with Hessian damping

$$\begin{aligned} x_{k+1} = x_k + \frac{1}{1+h\gamma}(x_k - x_{k-1}) - \frac{h\beta}{1+h\gamma}(\nabla f(x_k) - \nabla f(x_{k-1})) \\ - \frac{h^2}{1+h\gamma}\nabla f(x_k). \end{aligned} \tag{IGAHD}$$

(IGAHD) is based on the heavy ball with friction method. In the case of the accelerated gradient method of Nesterov, inertial algorithms involving a similar correcting term were studied recently by Attouch et al. [13] and Shi et al. [43]. We can now state the main results of the paper. In order not to make the statements too long, we expose separately the qualitative and the quantitative convergence results.

2.2. Convergence: Finite length property.

THEOREM 2.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ function whose gradient is L-Lipschitz continuous, and such that* $\inf_{\mathcal{H}} f > -\infty$. *Assume that the dry friction function ϕ*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

satisfies $(\mathrm{DF})_r$. *Suppose that the parameters* h, γ, β *in the algorithm* (IPAHDD) *satisfy the relation*

$$\gamma \geq L\left(\frac{h}{2}+\beta\right).$$

Then, for any sequence (x_k) *generated by the algorithm* (IPAHDD), *we have the following:*

(i) $\sum_{k=1}^{\infty}\|x_{k+1}-x_k\| < +\infty$, *and hence* $\lim_k x_k := x_\infty$ *exists for the strong topology of* $\mathcal{H}$. *Moreover,*

$$\sum_{k=1}^{\infty}\|x_{k+1}-x_k\| \leq \frac{1}{r}\left(E_1+\frac{\beta L}{2h}\|x_1-x_0\|^2\right),$$
$$\sum_{k=1}^{\infty}\|x_{k+1}-2x_k+x_{k-1}\|^2 \leq 2h^2\left(E_1+\frac{\beta L}{2h}\|x_1-x_0\|^2\right),$$

where $E_1 := \frac{1}{2}\|\frac{1}{h}(x_1-x_0)\|^2 + f(x_1) - \inf_{\mathcal{H}} f$.

(ii) *The limit* x_∞ *of the sequence* (x_k) *satisfies* $0 \in \partial\phi(0)+\nabla f(x_\infty)$.

Proof. We use an energetic argument based on the nonincreasing property of the sequence (E_k) of nonnegative global energy functions

$$E_k := \frac{1}{2}\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2 + f(x_k) - \inf_{\mathcal{H}} f.$$

Let's formulate (2.1) in terms of the discrete velocity vectors $\frac{1}{h}(x_k-x_{k-1})$. After multiplication by h, we obtain the equivalent formulation

$$\begin{aligned}&\frac{1}{h}(x_{k+1}-x_k)-\frac{1}{h}(x_k-x_{k-1})+\gamma(x_{k+1}-x_k)+h\partial\phi\left(\frac{1}{h}(x_{k+1}-x_k)\right)\\ &\quad+\beta\left(\nabla f(x_k)-\nabla f(x_{k-1})\right)+h\nabla f(x_k)\ni 0.\end{aligned} \tag{2.2}$$

(i) Let's first establish energy estimates. Without ambiguity, we write simply $\partial\phi$ to designate any element belonging to this set. Taking the scalar product of (2.2) with $\frac{1}{h}(x_{k+1}-x_k)$, we obtain

$$\begin{aligned}&\left\langle \frac{1}{h}(x_{k+1}-x_k)-\frac{1}{h}(x_k-x_{k-1}),\frac{1}{h}(x_{k+1}-x_k)\right\rangle+\gamma h\left\|\frac{1}{h}(x_{k+1}-x_k)\right\|^2\\ &\quad+\langle\nabla f(x_k),x_{k+1}-x_k\rangle+h\left\langle\partial\phi\left(\frac{1}{h}(x_{k+1}-x_k)\right),\frac{1}{h}(x_{k+1}-x_k)\right\rangle\\ &\quad+\beta\left\langle\nabla f(x_k)-\nabla f(x_{k-1}),\frac{1}{h}(x_{k+1}-x_k)\right\rangle=0.\end{aligned} \tag{2.3}$$

Set $X_k := \frac{1}{h}(x_k-x_{k-1})$. The following elementary relation reflects the strong convexity of $\frac{1}{2}\|\cdot\|^2$:

$$\langle X_{k+1}-X_k, X_{k+1}\rangle = \frac{1}{2}\|X_{k+1}\|^2-\frac{1}{2}\|X_k\|^2+\frac{1}{2}\|X_{k+1}-X_k\|^2. \tag{2.4}$$

According to the convexity of ϕ and $\phi(0)=0$, we have

$$\langle\partial\phi\left(X_{k+1}\right), X_{k+1}\rangle \geq \phi\left(X_{k+1}\right). \tag{2.5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Taking into account (2.4) and (2.5), we deduce from (2.3) the following inequality:

$$\frac{1}{2}\|X_{k+1}\|^2 - \frac{1}{2}\|X_k\|^2 + \frac{1}{2}\|X_{k+1} - X_k\|^2 + \gamma h\|X_{k+1}\|^2 + h\phi(X_{k+1}) \\ + \beta \left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1} - x_k) \right\rangle + \langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq 0. \tag{2.6}$$

Let's now use the assumptions on the potential functions ϕ and f. According to the assumption $(\mathrm{DF})_r$ on ϕ and Lemma 1.1, for all $k \geq 1$

$$\phi\left(X_{k+1}\right) \geq r\|X_{k+1}\|. \tag{2.7}$$

Since ∇f is L-Lipschitz continuous, the classical gradient descent lemma gives, for all $k \geq 1$,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{2.8}$$

According to the Cauchy–Schwarz inequality, and using again that ∇f is L-Lipschitz continuous,

$$\begin{aligned} \left|\left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1} - x_k) \right\rangle\right| &\leq hL\|X_k\|\|X_{k+1}\| \\ &\leq \frac{hL}{2}(\|X_k\|^2 + \|X_{k+1}\|^2). \end{aligned} \tag{2.9}$$

Combining inequalities (2.7)–(2.8)–(2.9) with (2.6), we obtain, for all $k \geq 1$,

$$\begin{aligned} &\frac{1}{2}\left\|\frac{1}{h}(x_{k+1} - x_k)\right\|^2 - \frac{1}{2}\left\|\frac{1}{h}(x_k - x_{k-1})\right\|^2 + \frac{1}{2h^2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 \\ &\qquad + \frac{\gamma}{h}\|x_{k+1} - x_k\|^2 + r\|x_{k+1} - x_k\| + f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq \frac{\beta L}{2h}(\|x_k - x_{k-1}\|^2 + \|x_{k+1} - x_k\|^2). \end{aligned}$$

In terms of $E_k := \frac{1}{2}\|\frac{1}{h}(x_k - x_{k-1})\|^2 + f(x_k) - \inf_{\mathcal{H}} f$, this is equivalent to

$$\begin{aligned} &E_{k+1} - E_k + \left(\frac{\gamma}{h} - \frac{L}{2} - \frac{\beta L}{2h}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{2h^2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 \\ &\qquad + r\|x_{k+1} - x_k\| \leq \frac{\beta L}{2h}\|x_k - x_{k-1}\|^2. \end{aligned}$$

According to the assumption $\gamma \geq L\left(\frac{h}{2} + \beta\right)$, we have $\frac{\gamma}{h} - \frac{L}{2} - \frac{\beta L}{2h} \geq \frac{\beta L}{2h}$. Therefore,

$$E_{k+1} - E_k + \frac{1}{2h^2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 + r\|x_{k+1} - x_k\| + \frac{\beta L}{2h}\|x_{k+1} - x_k\|^2 \leq \frac{\beta L}{2h}\|x_k - x_{k-1}\|^2. \tag{2.10}$$

Set $\tilde{E}_k := E_k + \frac{\beta L}{2h}\|x_k - x_{k-1}\|^2$. We have

$$\tilde{E}_{k+1} - \tilde{E}_k + \frac{1}{2h^2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 + r\|x_{k+1} - x_k\| \leq 0. \tag{2.11}$$

Adding the above inequalities, and according to $E_k \geq 0$, and $r > 0$, we deduce from (2.11) that

$$\sum_{k=1}^{\infty}\|x_{k+1} - x_k\| \leq \frac{1}{r}\left(E_1 + \frac{\beta L}{2h}\|x_1 - x_0\|^2\right) < +\infty. \tag{2.12}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Therefore, the sequence (x_k) has a finite length, which implies that the strong limit of the sequence (x_k) exists. Set $x_\infty := \lim_k x_k$. In addition, according to (2.11), we also get

$$\sum_{k=1}^{\infty} \|x_{k+1} - 2x_k + x_{k-1}\|^2 \leq 2h^2 \left(E_1 + \frac{\beta L}{2h} \|x_1 - x_0\|^2 \right) < +\infty. \tag{2.13}$$

Estimation (2.13) gives more accurate information than (2.12) when the step size h is small.

(ii) From $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < +\infty$, we get $\lim_k \|x_{k+1} - x_k\| = 0$. This in turn implies

$$\lim_k \frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) = \lim_k \frac{1}{h^2}\left((x_{k+1} - x_k) - (x_k - x_{k-1})\right) = 0.$$

Moreover, since ∇f is Lipschitz continuous and (x_k) converges strongly to x_∞, we have

$$\lim_k \nabla f(x_k) = \nabla f(x_\infty) \quad \text{and} \quad \lim_k \|\nabla f(x_k) - \nabla f(x_{k-1})\| = 0.$$

To pass to the limit on (2.1), rewrite it as follows:

$$\begin{aligned} &- \frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) - \frac{\gamma}{h}(x_{k+1} - x_k) - \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) \\ &- \nabla f(x_k) \in \partial\phi \left(\frac{1}{h}(x_{k+1} - x_k) \right). \end{aligned}$$

According to the above convergence results and the closedness of the graph of $\partial\phi$, we deduce that

$$-\nabla f(x_\infty) \in \partial\phi\,(0)\,,$$

which gives item (ii). ∎

2.3. Convergence rate: Geometric and finite convergence results. We have shown that the limit of the iterates x_∞ satisfies $-\nabla f(x_\infty) \in \partial\phi\,(0)$. We show that, when it happens that x_∞ satisfies the stronger property

$$- \nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0)), \tag{2.14}$$

we then obtain geometric convergence and finite convergence results. Note that condition (2.14) involves the limit of the iterates, namely x_∞, which is a priori unknown. But practically, this condition is almost always satisfied, making it a valuable numerical result.

THEOREM 2.2 (geometric, finite convergence). *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the dry friction function ϕ satisfies* $(\mathrm{DF})_r$. *Suppose that the positive parameters h, γ, β in* (IPAHDD) *satisfy the relation*

$$\gamma \geq L\left(\frac{h}{2} + \beta\right).$$

Let (x_k) be a sequence generated by the algorithm (IPAHDD), *and let x_∞ be its strong limit (as given by Theorem* 2.1*).*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

(i) *Suppose that* $-\nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0))$. *Then, there is geometric convergence of the velocities to zero. Set* $q := \frac{1}{\sqrt{1+\frac{2h(\gamma-\beta L)}{1+\beta hL}}}$, *which satisfies* $0 < q < 1$. *There exists* $k_0 \geq 0$ *such that*

$$\forall k \geq k_0 \quad \|x_{k+1} - x_k\| \leq q^k \|x_{k_0+1} - x_{k_0}\|.$$

There is geometric convergence of the sequence (x_k): *for all* $k \geq k_0$

$$\|x_k - x_\infty\| \leq \frac{q^k}{1-q}\|x_{k_0+1} - x_{k_0}\|. \tag{2.15}$$

(ii) *Suppose that* $\|\nabla f(x_\infty)\| < r$, *where* $B(0,r) \subset \partial\phi(0)$. *Then the sequence* (x_k) *is finitely convergent. The iteration stops at* x_k *when* $k \geq k_0$ *and*

$$q^{k-1} \leq \frac{r - \|\nabla f(x_\infty)\|}{\left(\frac{1}{h^2} + \frac{\beta L}{h} + L\frac{q}{1-q}\right)\|x_{k_0+1} - x_{k_0}\|},$$

which is satisfied for k *large enough, because of* $q < 1$.

Proof. (i) The assumption $-\nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0))$ implies the existence of $\varepsilon > 0$ such that

$$-\nabla f(x_\infty) + B(0, 2\varepsilon) \subset \partial\phi(0).$$

On the other hand, since $\lim_k \nabla f(x_k) = \nabla f(x_\infty)$, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$

$$\nabla f(x_k) \in \nabla f(x_\infty) + B(0,\varepsilon).$$

Therefore,

$$-\nabla f(x_k) + B(0,\varepsilon) \subset -\nabla f(x_\infty) + B(0, 2\varepsilon) \subset \partial\phi(0).$$

Equivalently, for every $k \geq k_0$ and for every $u \in B(0,1)$, we have $-\nabla f(x_k) + \varepsilon u \in \partial\phi(0)$.

Let's write the corresponding subdifferential inequality at the origin (recall that $\phi(0) = 0$). For every $k \geq k_0$, we have

$$\forall u \in B(0,1), \quad \phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) \geq \left\langle -\nabla f(x_k) + \varepsilon u, \frac{1}{h}(x_{k+1} - x_k)\right\rangle.$$

Taking the supremum over $u \in B(0,1)$, we obtain that, for every $k \geq k_0$,

$$\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) + \left\langle \nabla f(x_k), \frac{1}{h}(x_{k+1} - x_k)\right\rangle \geq \varepsilon \left\|\frac{1}{h}(x_{k+1} - x_k)\right\|. \tag{2.16}$$

Let's return to inequality (2.6). According to (2.9), we have

$$\frac{1}{2}\left\|\frac{1}{h}(x_{k+1} - x_k)\right\|^2 - \frac{1}{2}\left\|\frac{1}{h}(x_k - x_{k-1})\right\|^2 + \left(\frac{\gamma}{h} - \frac{\beta L}{2h}\right)\|x_{k+1} - x_k\|^2$$
$$+ h\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq \frac{\beta L}{2h}\|x_k - x_{k-1}\|^2.$$

According to $\gamma \geq L(\frac{h}{2} + \beta)$, we have $\gamma > \beta L$. Hence $\frac{\gamma}{h} - \frac{\beta L}{2h} = \frac{\beta L}{2h} + \frac{1}{h}(\gamma - \beta L) > \frac{\beta L}{2h}$, which gives

$$\frac{1}{2}\left\|\frac{1}{h}(x_{k+1} - x_k)\right\|^2 - \frac{1}{2}\left\|\frac{1}{h}(x_k - x_{k-1})\right\|^2 + \frac{\beta L}{2h}\|x_{k+1} - x_k\|^2 + \frac{1}{h}(\gamma - \beta L)\|x_{k+1} - x_k\|^2$$
$$+ h\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq \frac{\beta L}{2h}\|x_k - x_{k-1}\|^2.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Combining the inequality above with (2.16), we obtain, for every $k \geq k_0$,

$$\begin{aligned}&\frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_{k+1}-x_k)\right\|^2-\frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2\\ &\quad+\frac{1}{h}\left(\gamma-\beta L\right)\|x_{k+1}-x_k\|^2+\varepsilon\,\|x_{k+1}-x_k\|\leq 0.\end{aligned} \tag{2.17}$$

Neglecting the nonnegative term $\varepsilon\,\|x_{k+1}-x_k\| \geq 0$ in the above inequality, we obtain

$$\frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_{k+1}-x_k)\right\|^2-\frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2+\frac{1}{h}\left(\gamma-\beta L\right)\|x_{k+1}-x_k\|^2\leq 0.$$

Equivalently

$$\Big(1+\beta hL+2h(\gamma-\beta L)\Big)\|x_{k+1}-x_k\|^2\leq\Big(1+\beta hL\Big)\|x_k-x_{k-1}\|^2,$$

which gives the geometric convergence of velocities toward zero: for $k \geq k_0$

$$\|x_{k+1}-x_k\|\leq q^k\|x_{k_0+1}-x_{k_0}\|, \tag{2.18}$$

with $q := \frac{1}{\sqrt{1+\frac{2h(\gamma-\beta L)}{1+\beta hL}}}$. Set $C := \|x_1 - x_0\|$. For, $p \geq 0$ we have

$$\|x_k-x_{k+p}\|\leq Cq^k\left(1+q+\cdots+q^{p-1}\right)\leq C\frac{q^k}{1-q}.$$

By making p go to infinity in the inequality above, and using that (x_k) converges to x_∞, we obtain

$$\|x_k-x_\infty\|\leq C\frac{q^k}{1-q}.$$

This formula expresses the geometric convergence of the sequence (x_k) to its limit x_∞. This is a remarkable property because there can be a continuum of possible limits of the sequence (x_k).

(ii) Let us show that the finite convergence property holds under the assumption $\|\nabla f(x_\infty)\| < r$ where $B(0,r) \subset \partial\phi(0)$. Write the algorithm (IPAHDD) as follows:

$$\frac{1}{h}(x_{k+1}-x_k)+\gamma(x_{k+1}-x_k)+h\partial\phi\left(\frac{1}{h}(x_{k+1}-x_k)\right)\ni\xi_k, \tag{2.19}$$

where $\xi_k := \frac{1}{h}(x_k - x_{k-1}) - h\nabla f(x_k) - \beta\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)$. Equivalently,

$$(1+\gamma h)\left(\frac{1}{h}(x_{k+1}-x_k)\right)+h\partial\phi\left(\frac{1}{h}(x_{k+1}-x_k)\right)\ni\xi_k,$$

which gives, with $\lambda := \frac{h}{1+\gamma h}$,

$$\frac{1}{h}(x_{k+1}-x_k)=\operatorname{prox}_{\lambda\phi}\left(\frac{1}{1+\gamma h}\xi_k\right). \tag{2.20}$$

To show the finite convergence property, we need to show that $x_{k+1} - x_k = 0$ for k large enough. According to (2.20) and Lemma 1.2, it suffices to prove that

$$\frac{1}{\lambda}\left\|\frac{1}{1+\gamma h}\xi_k\right\|\leq r, \tag{2.21}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

which, according to the definition of λ, is written equivalently as $\frac{1}{h}\|\xi_k\| \le r$. By the triangle inequality and the L-Lipschitz continuity of ∇f we have

$$\frac{1}{h}\|\xi_k\| = \left\| \frac{1}{h^2}(x_k - x_{k-1}) - \nabla f(x_k) - \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)\right\|$$
$$\le \frac{1}{h^2}\|x_k - x_{k-1}\| + \|\nabla f(x_\infty)\| + L\|x_k - x_\infty\| + \frac{\beta L}{h}\|x_k - x_{k-1}\|. \tag{2.22}$$

When $k \to +\infty$, the right-hand side of the inequality (2.22) tends to $\|\nabla f(x_\infty)\|$. So, condition (2.21) will be satisfied for k large enough if $\|\nabla f(x_\infty)\| < r$. Let us suppose this condition is satisfied and further analyze (2.22). We have $x_{k+1} - x_k = 0$ as soon as

$$\left(\frac{1}{h^2} + \frac{\beta L}{h}\right)\|x_k - x_{k-1}\| + L\|x_k - x_\infty\| \le r - \|\nabla f(x_\infty)\|.$$

According to the geometric convergence rate obtained in (i), this will be satisfied when $k \ge k_0$ and

$$\left(\frac{1}{h^2} + \frac{\beta L}{h} + L\frac{q}{1-q}\right) q^{k-1}\|x_{k_0+1} - x_{k_0}\| \le r - \|\nabla f(x_\infty)\|.$$

This gives

$$q^{k-1} \le \frac{r - \|\nabla f(x_\infty)\|}{\left(\frac{1}{h^2} + \frac{\beta L}{h} + L\frac{q}{1-q}\right)\|x_{k_0+1} - x_{k_0}\|},$$

which completes the proof. ∎

Remark 2.3. Let's give another proof of the finite convergence property. On the one hand, it only assumes that $-\nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0))$, but it is valid only when $\mathcal{H}$ is a finite dimensional space. It is similar to the argument developed by Baji and Cabot in [22]. Argue by contradiction, and suppose that there is an infinite number of indices $k \in \mathbb{N}$ such that $\|x_{k+1} - x_k\| \neq 0$. Set $\mathcal{N} := \{k \in \mathbb{N} : \|x_{k+1} - x_k\| \neq 0\}$, and consider the sequence $(\omega_k)_k$ defined by

$$\omega_k := \frac{x_{k+1} - x_k}{\|x_{k+1} - x_k\|} \quad \text{for} \quad k \in \mathcal{N}.$$

The sequence (ω_k) belongs to the unit sphere of $\mathcal{H}$, and since $\mathcal{H}$ is assumed to have a finite dimension, we can extract a convergent sequence (still denoted (ω_k)) that converges to a point ω which belongs to the unit sphere (in an infinite dimensional space, we would only have weak convergence toward a point of the unit ball). Set $a_k = -\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1})$. According to the monotonicity property of $\partial\phi$ and the definition of the algorithm (IPAHDD), we have, for all $k \in \mathcal{N}$,

$$\left\langle a_k - \frac{\gamma}{h}(x_{k+1} - x_k) - \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \nabla f(x_k) - \partial\phi(0), \omega_k \right\rangle \ge 0, \tag{2.23}$$

where, in the above formula, $\partial\phi(0)$ denotes an arbitrary point of this set.

According to convergence properties shown above, by passing to the limit in (2.23), we obtain

$$\langle \nabla f(x_\infty) + \partial\phi(0), \omega\rangle \le 0.$$

Since $-\nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0))$, there exists some $\rho > 0$ such that $B(0,\rho) \subset \nabla f(x_\infty) + \partial\phi(0)$.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Therefore, we would have $\langle \rho u, \omega \rangle \leq 0$ for all $u \in B(0,1)$. Taking $u = \omega$ (since $\|\omega\| = 1$) gives $\rho \|\omega\|^2 \leq 0$, and hence $\omega = 0$, a clear contraction with ω belonging to the unit sphere.

Remark 2.4. The case $\phi = 0$ gives the heavy ball with friction method initiated by Polyak [41], [42]. This case is excluded from our analysis because of the dry friction hypothesis $(\mathrm{DF})_r$ on ϕ. We can advantageously compare our method with the heavy ball method, for which the results of convergence require restrictive assumptions on the parameters and the function f; see [33] for a recent account on the heavy ball method. Note that, compared to the restart method, we get a geometric convergence for a general function f, which might be nonconvex.

2.4. Soft thresholding on the velocities. As a model situation for dry friction, take $\phi : \mathcal{H} \to \mathbb{R}$ given by $\phi(x) = r\|x\|$, with $r > 0$. We have

$$\partial \phi(x) = \begin{cases} r \frac{x}{\|x\|} & \text{if } x \neq 0; \\ B(0,r) & \text{if } x = 0. \end{cases} \tag{2.24}$$

By definition of the proximal operator, we obtain, for all $\lambda > 0$,

$$\operatorname{prox}_{\lambda \phi}(x) = \left(1 - \frac{\lambda r}{\max\{\lambda r, \|x\|\}}\right) x = \begin{cases} 0 & \text{if } \|x\| \leq \lambda r; \\ (\|x\| - \lambda r)\, \frac{x}{\|x\|} & \text{if } \|x\| \geq \lambda r. \end{cases} \tag{2.25}$$

(a) When $\mathcal{H} = \mathbb{R}$, we get the classical soft thresholding operator $\operatorname{prox}_{\lambda\phi} = T_{\lambda r}$, which is used in the FISTA method for sparse optimization:

$$T_{\lambda r}(x) = \operatorname{sign}(x)(|x| - \lambda r)_+ = \begin{cases} x - \lambda r & \text{if } x \geq \lambda r; \\ 0 & \text{if } -\lambda r \leq x \leq \lambda r; \\ x + \lambda r & \text{if } x \leq -\lambda r. \end{cases} \tag{2.26}$$

(b) In the multidimensional case $\mathcal{H} = \mathbb{R}^n$, take $\phi : \mathbb{R}^n \to \mathbb{R}$ given by $\phi(x) = r\|x\|_1 = r\sum_{i=1}^n |x_i|$. The proximal mapping of ϕ can be computed componentwise by applying the one-dimensional soft thresholding operator $T_{\lambda r}$ to each component. This is transparent from the variational formulation of the proximal operator: $\operatorname{prox}_{\lambda\phi}(x)$ is the solution of the minimization problem

$$\min_{\xi \in \mathbb{R}^n} \left\{ \frac{1}{2}\|\xi - x\|^2 + \lambda r \|\xi\|_1 \right\} = \min_{\xi_1 \in \mathbb{R}, \dots, \xi_n \in \mathbb{R}} \left\{ \sum_i \left(\frac{1}{2}|\xi_i - x_i|^2 + \lambda r |\xi_i| \right) \right\}$$

which can be decomposed with respect to each component. Hence

$$\left(\operatorname{prox}_{\lambda r \|\cdot\|_1}(x)\right)_i = T_{\lambda r}(x_i) = \operatorname{sign}(x_i)(|x_i| - \lambda r)_+ \quad \text{for } i = 1, 2, \dots n. \tag{2.27}$$

The algorithm (IPAHDD) is a splitting algorithm which reads componentwise as follows: setting $x_k = (x_{k,i})_{i=1,2,\dots n}$, we have for $i = 1, 2, \dots n$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

(IPAHDD) with soft thresholding on the velocities

Initialize: $x_0 \in \mathbb{R}^n$, $x_1 \in \mathbb{R}^n$

$for\ i = 1, 2, \dots n$

$$x_{k+1,i} = x_{k,i} + hT_{\frac{hr}{1+h\gamma}}\left(\frac{1}{h(1+h\gamma)}(x_{k,i} - x_{k-1,i}) - \frac{\beta}{1+h\gamma}\left(\frac{\partial f}{\partial x_i}(x_k) - \frac{\partial f}{\partial x_i}(x_{k-1})\right) - \frac{h}{1+h\gamma}\frac{\partial f}{\partial x_i}(x_k)\right)$$

The operator $T_{\frac{hr}{1+h\gamma}}$ acts as a soft thresholding operator on the velocities. A direct application of Theorems 2.1 and 2.2 gives the following result.

COROLLARY 2.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$ function whose gradient is L-Lipschitz continuous, and such that* $\inf_{\mathcal{H}} f > -\infty$. *Assume that the potential friction function ϕ is given by $\phi(x) = r\|x\|_1$. Suppose that the positive parameters h, γ, β in the algorithm* (IPAHDD) *satisfy the relation*

$$\gamma \geq L\left(\frac{h}{2} + \beta\right).$$

Let (x_k) be a sequence generated by the algorithm (IPAHDD) *with soft thresholding on the velocities.*

(i) *Then, the sequence (x_k) has a finite length. It converges to x_∞ that verifies $\|\nabla f(x_\infty)\|_1 \leq r$.*

(ii) *Suppose that $\|\nabla f(x_\infty)\|_1 < r$. Then, there is geometric convergence* (2.15), *and the sequence (x_k) is finitely convergent.*

Clearly, taking r small is the interesting situation for optimization.

2.5. An example. Take $\mathcal{H} = \mathbb{R}$, $\phi(x) = r|x|$, $r > 0$, and $f(x) = \frac{1}{2}x^2$. With $h = 1$, the algorithm (IPAHDD) reads as follows:

$$(x_{k+1}-x_k)-(x_k-x_{k-1})+\gamma(x_{k+1}-x_k)+\beta(x_k-x_{k-1})+\partial\phi(x_{k+1} - x_k)+x_k \ni 0. \tag{2.28}$$

Equivalently, $(x_{k+1}-x_k)+\frac{1}{1+\gamma}\partial\phi(x_{k+1} - x_k) \ni -\frac{1}{1+\gamma}(x_{k-1} + \beta(x_k - x_{k-1}))$, which gives

$$x_{k+1} - x_k = T_{\frac{r}{1+\gamma}}\left(-\frac{1}{1+\gamma}x_{k-1} - \frac{\beta}{1+\gamma}(x_k - x_{k-1})\right). \tag{2.29}$$

According to (2.26), with $\lambda = \frac{1}{1+\gamma}$, we obtain

$$x_{k+1}-x_k = \begin{cases} -\dfrac{1}{1+\gamma}(x_{k-1}+r) - \dfrac{\beta}{1+\gamma}(x_k - x_{k-1}) & \text{if } x_{k-1} + \beta(x_k - x_{k-1}) \leq -r; \\ 0 & \text{if } |x_{k-1} + \beta(x_k - x_{k-1})| \leq r; \\ -\dfrac{1}{1+\gamma}(x_{k-1}-r) - \dfrac{\beta}{1+\gamma}(x_k - x_{k-1}) & \text{if } x_{k-1} + \beta(x_k - x_{k-1}) \geq r. \end{cases}$$

Take $\gamma = 3$, $\beta = 1$. We have $L = 1$, and the condition $\gamma \geq L\left(\frac{h}{2} + \beta\right)$ of Theorem 2.2 is satisfied. So, as long as $x_k \geq r$, according to the above formula, we have

$$x_{k+1} - x_k = -\frac{1}{1+\gamma}(x_{k-1} - r) - \frac{\beta}{1+\gamma}(x_k - x_{k-1}) = -\frac{1}{4}(x_k - r).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

The sequence $(x_k - r)$ satisfies the geometric recurrence relation $x_{k+1} - 1 = \frac{3}{4}(x_k - r)$, which gives

$$x_k = r + \left(\frac{3}{4}\right)^{k-1}(x_1 - r).$$

By taking $x_1 \geq r$, the condition $x_k \geq r$ is satisfied. So, in this particular situation we have linear convergence but not finite convergence. This is in accordance with the fact that $x_\infty = r$ and that $\nabla f(x_\infty) = r$, which is not in the interior of the convex set $\partial\phi(0) = [-r, +r]$.

2.6. A variant. Consider the following discretization of the differential inclusion (IGDH):

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_k - x_{k-1}) + \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right)$$
$$+ \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) + \nabla f(x_k) \ni 0,$$

where the temporal discretization of the viscous damping term is taken equal to $\frac{\gamma}{h}(x_k - x_{k-1})$ instead of $\frac{\gamma}{h}(x_{k+1} - x_k)$. Solving this equation with respect to x_{k+1} gives the following algorithm:

(IPAHDD-Var)

Initialize: $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$

$x_{k+1} = x_k + h\,\mathrm{prox}_{h\phi}\Big(\Big(\frac{1-h\gamma}{h}\Big)(x_k - x_{k-1}) - \beta\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - h\nabla f(x_k)\Big)$

The proof of the following convergence result is similar to Theorems 2.1–2.2 and hence is omitted.

Theorem 2.6. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the dry friction function ϕ satisfies* $(\mathrm{DF})_r$. *Suppose that the positive parameters h, γ in the algorithm* (IPAHDD-Var) *satisfy the relation*

$$\gamma \geq L\left(\beta + \frac{1}{2}h\right) + \frac{1}{2}\gamma^2 h.$$

Then, for any sequence (x_k) defined by the algorithm (IPAHDD-Var), *we have the following:*

(i) *$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < +\infty$, and $\lim_k x_k := x_\infty$ exists for the strong topology of $\mathcal{H}$. Moreover,*

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \leq \frac{E_1}{r}, \ \textit{where}\ E_1 := \frac{1}{2}\Big(1 + \beta h L\Big)\left\|\frac{1}{h}(x_1 - x_0)\right\|^2 + f(x_1) - \inf_{\mathcal{H}} f.$$

(ii) *The limit x_∞ of the sequence (x_k) satisfies $0 \in \partial\phi(0) + \nabla f(x_\infty)$.*

(iii) *Suppose that $\gamma > L\left(\beta + \frac{1}{2}h\right) + \frac{1}{2}\gamma^2 h$.*

  (a) *Suppose that $-\nabla f(x_\infty) \in \mathrm{int}(\partial\phi(0))$. Then, there is geometric convergence of (x_k).*

  (b) *Suppose that $\|\nabla f(x_\infty)\| < r$, where $B(0, r) \subset \partial\phi(0)$. Then (x_k) is finitely convergent.*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

2.7. Errors, perturbations. Let's introduce perturbations and errors in the algorithm (IPAHDD). According to the dynamic approach, we start from the perturbed version of (IGDH)

$$\ddot{x}(t) + \gamma \dot{x}(t) + \partial\phi(\dot{x}(t)) + \beta \nabla^2 f(\dot{x}(t)) + \nabla f(x(t)) \ni e(t), \tag{2.30}$$

where the second member $e(\cdot)$ takes into account perturbations and errors. A similar temporal discretization as in section 2 gives

$$\begin{aligned} &\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) \\ &+\frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) + \nabla f(x_k) \ni e_k. \end{aligned} \tag{2.31}$$

Solving (2.31) with respect to x_{k+1} gives the following algorithm:

(IPAHDD-pert)

Initialize: $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$

$y_k = \frac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \frac{\beta}{1+h\gamma}(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{h}{1+h\gamma}\nabla f(x_k) + \frac{h}{1+h\gamma} e_k$

$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}(y_k)$

The following convergence results are parallel to Theorems 2.1 and 2.2. Their proof is given in the appendix.

THEOREM 2.7. *Let's make the assumptions of Theorem* 2.1, *and suppose that the sequence* (e_k) *of perturbations and errors satisfies* $\lim_k \|e_k\| = 0$. *Then, for any sequence* (x_k) *defined by the algorithm* (IPAHDD-pert), *we have the following:*

(i) $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < +\infty$, *and therefore* $\lim_k x_k := x_\infty$ *exists for the strong topology of* $\mathcal{H}$.
Suppose that $\|e_k\| \le \frac{r}{2}$. *Then*

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \le \frac{2}{r}\left(E_1 + \frac{\beta L}{2h}\|x_1 - x_0\|^2\right).$$

(ii) *The limit* x_∞ *of the sequence* (x_k) *satisfies* $0 \in \partial\phi(0) + \nabla f(x_\infty)$.

(iii) *Suppose that* $-\nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0))$. *Then, there is geometric convergence of the velocities to zero. Set* $q = \frac{1}{\sqrt{1+2h\gamma}}$. *There exists* $k_0 \ge 0$ *such that for all* $k \ge k_0$

$$\|x_k - x_\infty\| \le \frac{q^k}{1-q}\|x_{k_0+1} - x_{k_0}\|.$$

(iv) *Suppose that* $\|\nabla f(x_\infty)\| < r$, *where* $B(0, r) \subset \partial\phi(0)$. *Then* (x_k) *is finitely convergent.*

Remark 2.8. An interesting question would be to study the case of relative errors; see [38] for a recent account on the subject. Note that most of the existing results concern the case of convex optimization or monotone inclusions, while our study works without any convexity assumption on the function f to minimize.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

3. Combining Nesterov acceleration method with dry friction. We construct algorithms, still obtained by temporal discretizations of the differential inclusion

$$\text{(IGDH)} \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \partial\phi(\dot{x}(t)) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) \ni 0,$$

and which have an analogous structure to the accelerated gradient method of Nesterov [39, 40]. Indeed, when discretizing (IGDH), there is some flexibility in the choice of the point y_k where the gradient of f is computed. By taking $y_k = x_k$, we obtain the algorithm (IPAHDD) studied in the previous section. We consider two different choices for y_k, which are in accordance with the Nesterov accelerated gradient method. They lead respectively to the so-called (IPAHDD-N) and (IPAHDD-N-Var) algorithms, where the suffix "N" refers to Nesterov and the suffix "Var" refers to variant. Since their asymptotic convergence analysis is based on an argument similar to that used for (IPAHDD), their proofs are given in the appendix. While from a theoretical point of view these algorithms can be considered as variants of (IPAHDD), we observe in section 5.2 the importance and empirical success of (IPAHDD-N-Var).

3.1. (IPAHDD-N). First, let us consider the following temporal discretization of (IGDH):

$$\begin{aligned}\text{(3.1)} \qquad &\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) \\ &\quad + \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) + \nabla f(y_k) \ni 0,\end{aligned}$$

where y_k will be chosen consistently with the accelerated gradient method of Nesterov. To solve (3.1) with respect to $\frac{1}{h}(x_{k+1} - x_k)$, let's write it equivalently as

$$\begin{aligned}&\frac{1}{h}(x_{k+1} - x_k) - \frac{1}{h}(x_k - x_{k-1}) + h\gamma\frac{1}{h}(x_{k+1} - x_k) + h\partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) \\ \text{(3.2)} \quad &+ \beta\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) + h\nabla f(y_k) \ni 0.\end{aligned}$$

Equivalently,

$$\frac{1}{h}(x_{k+1} - x_k) + \frac{h}{1+h\gamma}\partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right) \ni z_k,$$

with $z_k := \frac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \frac{\beta}{1+h\gamma}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{h}{1+h\gamma}\nabla f(y_k)$. Therefore,

$$\text{(3.3)} \qquad x_{k+1} = x_k + h\,\mathrm{prox}_{\frac{h}{1+h\gamma}\phi}(z_k).$$

When $\phi = 0$, and $\beta = 0$, the proximal operator is the identity, and we obtain

$$x_{k+1} = x_k + \frac{1}{1+h\gamma}(x_k - x_{k-1}) - \frac{h^2}{1+h\gamma}\nabla f(y_k).$$

To recover the accelerated gradient method of Nesterov, we must take $y_k = x_k + \frac{1}{1+h\gamma}(x_k - x_{k-1})$. In doing so, we obtain the following algorithm:

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

(IPAHDD-N)

Initialize: $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$

$y_k = x_k + \frac{1}{1+h\gamma}(x_k - x_{k-1})$

$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}\left(\frac{1}{h}(y_k - x_k) - \frac{\beta}{1+h\gamma}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{h}{1+h\gamma}\nabla f(y_k)\right)$

The convergence properties of (IPAHDD-N) are analyzed in the following theorem (its proof is postponed to Appendix B). Note that the two sequences (x_k) and (y_k) enjoy similar convergence properties.

THEOREM 3.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ function whose gradient is L-Lipschitz continuous and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the potential friction function ϕ satisfies* $(\mathrm{DF})_r$. *Suppose that the positive parameters h, γ, β in the algorithm* (IPAHDD-N) *satisfy the relation*

$$\gamma \geq \frac{3L}{2}(h+\beta) \quad \textit{and} \quad Lh^2 \leq 1.$$

Then, for any sequence (x_k) defined by the algorithm (IPAHDD-N)*, we have the following:*

(i) $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < \infty$*, and hence $\lim_k x_k := x_\infty$ exists for the strong topology of $\mathcal{H}$. Moreover,*

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \leq \frac{1}{r}\left(\frac{1}{2h^2}\left(1 + h\gamma - \frac{Lh^2}{2}\right)\|x_1 - x_0\|^2 + f(x_1) - \inf_{\mathcal{H}} f\right).$$

(ii) *The limit x_∞ of the sequence (x_k) satisfies $0 \in \partial\phi(0) + \nabla f(x_\infty)$.*

(iii) *Suppose that $-\nabla f(x_\infty) \in \operatorname{int}(\partial\phi(0))$. Then, there is geometric convergence of (x_k).*

(iv) *Suppose that $\|\nabla f(x_\infty)\| < r$, where $B(0,r) \subset \partial\phi(0)$. Then (x_k) is finitely convergent.*

Similar properties hold for the sequence (y_k), which has the same limit as (x_k).

3.2. (IPAHDD-N-Var). Let's go back to (3.3), which is recalled here,

$$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}(z_k),$$

with $z_k = \frac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \frac{\beta}{1+h\gamma}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{h}{1+h\gamma}\nabla f(y_k)$, and make a different choice of the extrapolated point y_k. Taking $y_k - x_k = \frac{1}{h(1+h\gamma)}(x_k - x_{k-1})$ gives the following algorithm:

(IPAHDD-N-Var)

Initialize: $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$

$y_k = x_k + \frac{1}{h(1+h\gamma)}(x_k - x_{k-1})$

$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}\left(y_k - x_k - \frac{\beta}{1+h\gamma}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{h}{1+h\gamma}\nabla f(y_k)\right)$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

When $\phi = 0$ and $\beta = 0$, we obtain

$$x_{k+1} = x_k + \frac{1}{1+h\gamma}(x_k - x_{k-1}) - \frac{h^2}{1+h\gamma}\nabla f\left(x_k + \frac{1}{h(1+h\gamma)}(x_k - x_{k-1})\right).$$

This corresponds to a variant of the Nesterov accelerated gradient method, with two different extrapolation coefficients $\alpha_{k,1} = \frac{1}{1+h\gamma}$ and $\alpha_{k,2} = \frac{1}{h(1+h\gamma)}$. This type of situation has been studied by Liang, Fadili, and Peyré in [36]. In a recent paper [34], László obtained convergence rates for this method, showing that taking different extrapolation coefficients could be judicious. Understanding the best adjustment of the two extrapolation coefficients is a current research subject. From a theoretical point of view (IPAHDD-N-Var) has similar convergence properties to (IPAHDD-N). However, according to the comments above, we notice in section 5.2 that (IPAHDD-N-Var) performs better numerically.

4. (IPAHDD) for nonsmooth functions. We assume that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semicontinuous and proper function such that $\inf_{\mathcal{H}} f > -\infty$. The preceding sections deal with a differentiable function f, without convexity assumption on f. Now, when considering nonsmooth functions, we assume the convexity of f. This allows us to use the regularity properties of the Moreau envelope in the convex case. Indeed, to reduce to the previous situation, where $f : \mathcal{H} \to \mathbb{R}$ is a $\mathcal{C}^1$ function whose gradient is Lipschitz continuous, the idea is to replace f by its Moreau envelope. Recall some classical facts. For any $\lambda > 0$, the Moreau envelope of f of index λ is the function $f_\lambda : \mathcal{H} \to \mathbb{R}$ defined by, for all $x \in \mathcal{H}$,

$$f_\lambda(x) = \min_{\xi \in \mathcal{H}} \left\{ f(\xi) + \frac{1}{2\lambda}\|x - \xi\|^2 \right\}.$$

The function f_λ is convex, of class $\mathcal{C}^{1,1}$, and such that $\inf_{\mathcal{H}} f_\lambda = \inf_{\mathcal{H}} f$, $\operatorname{argmin}_{\mathcal{H}} f_\lambda = \operatorname{argmin}_{\mathcal{H}} f$. One can consult [9, section 17.2.1], [23], [29] for an in-depth study of the properties of the Moreau envelope in a Hilbert framework. Since the infimal value and the set of minimizers are preserved by taking the Moreau envelope, the idea is to replace f by f_λ in the previous algorithm and take advantage of the fact that f_λ is continuously differentiable. The algorithm (IPAHDD) becomes

$$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}(y_k)$$

with

$$y_k = \frac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \frac{\beta}{1+h\gamma}(\nabla f_\lambda(x_k) - \nabla f_\lambda(x_{k-1})) - \frac{h}{1+h\gamma}\nabla f_\lambda(x_k).$$

According to $\nabla f_\lambda(x) = \frac{1}{\lambda}\left(x - \operatorname{prox}_{\lambda f}(x)\right)$, we obtain the following:

(IPAHDD-nonsmooth)

Initialize: $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$

$y_k = \frac{1}{1+h\gamma}\left(\frac{\lambda - \beta h}{\lambda h}(x_k - x_{k-1}) + \frac{\beta}{\lambda}\left(\operatorname{prox}_{\lambda f}(x_k) - \operatorname{prox}_{\lambda f}(x_{k-1})\right) - \frac{h}{\lambda}\left(x_k - \operatorname{prox}_{\lambda f}(x_k)\right)\right)$

$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}(y_k)$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Note that the two nonsmooth functions f and ϕ enter the algorithm via their proximal mappings. In addition, these proximal steps are computed independently, which makes (IPAHDD-nonsmooth) a splitting algorithm. Based on the properties of the Moreau envelope, a direct adaptation of Theorem 2.1 gives the following convergence results for algorithm (IPAHDD-nonsmooth).

THEOREM 4.1. *Let* $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *be a convex lower semicontinuous and proper function such that* $\inf f > -\infty$. *Assume that the potential friction function* ϕ *satisfies* $(\mathrm{DF})_r$. *Suppose that the parameters* h, γ, β, λ *in the algorithm* (IPAHDD-nonsmooth) *satisfy the relation*

$$\gamma \geq \frac{1}{\lambda}\left(\frac{h}{2} + \beta\right).$$

Then, for any sequence (x_k) *defined by the algorithm* (IPAHDD-nonsmooth)*, we have the following:*

(i) $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < +\infty$. *Hence* $\lim_k x_k := x_\infty$ *exists for the strong topology of* $\mathcal{H}$. *Moreover,*

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \leq \frac{1}{r}\left(E_1 + \frac{\beta L}{2h}\|x_1 - x_0\|^2\right),$$

$$\sum_{k=1}^{\infty} \|x_{k+1} - 2x_k + x_{k-1}\|^2 \leq 2h^2\left(E_1 + \frac{\beta L}{2h}\|x_1 - x_0\|^2\right),$$

where $E_1 := \frac{1}{2}\|\frac{1}{h}(x_1 - x_0)\|^2 + f(x_1) - \inf_{\mathcal{H}} f$.

(ii) *The limit* x_∞ *of the sequence* (x_k) *satisfies* $0 \in \partial\phi(0) + \nabla f_\lambda(x_\infty)$. *The sequence* (p_k) *with* $p_k = \mathrm{prox}_{\lambda f}(x_k)$ *has a finite length, and it converges strongly toward* $p_\infty = \mathrm{prox}_{\lambda f}\, x_\infty$ *which satisfies the approximate optimality property:*

$$\partial f(p_\infty) + \partial\phi(0) \ni 0.$$

Suppose moreover that $-\nabla f_\lambda(x_\infty) \in \mathrm{int}(\partial\phi(0))$. *Then, there is geometric convergence of the velocities to zero. Set* $q := \frac{1}{\sqrt{1+\frac{2h(\gamma\lambda-\beta)}{\lambda+\beta h}}} \in]0,1[$. *There exists* $k_0 \geq 0$ *such that for all* $k \geq k_0$

$$\|x_{k+1} - x_k\| \leq q^k \|x_{k_0+1} - x_{k_0}\| \quad \text{and} \quad \|x_k - x_\infty\| \leq \frac{q^k}{1-q}\|x_{k_0+1} - x_{k_0}\|.$$

(iii) *Suppose that* $\|\nabla f_\lambda(x_\infty)\| < r$, *where* $B(0,r) \subset \partial\phi(0)$. *Then, the sequence* (x_k) *is finitely convergent. The iteration stops at* x_k *when* $k \geq k_0$ *and*

$$q^{k-1} \leq \frac{r - \|\nabla f_\lambda(x_\infty)\|}{\left(\frac{1}{h^2} + \frac{\beta}{h\lambda} + \frac{q}{\lambda(1-q)}\right)\|x_{k_0+1} - x_{k_0}\|},$$

which is satisfied for k *large enough, because of* $q < 1$.

Proof. The proof consists in replacing f by f_λ in Theorems 2.1 and 2.2 and in using that ∇f_λ is $\frac{1}{\lambda}$-Lipschitz continuous. By taking $L = \frac{1}{\lambda}$, the condition $\gamma \geq L(\frac{h}{2} + \beta)$ becomes $\gamma \geq \frac{1}{\lambda}(\frac{h}{2} + \beta)$. Now consider the sequence (p_k) with $p_k = \mathrm{prox}_{\lambda f}\, x_k$. Since $\mathrm{prox}_{\lambda f}$ is a nonexpansive mapping, we have $\sum_k \|p_{k+1} - p_k\| \leq \sum_k \|x_{k+1} - x_k\| < +\infty$. Consequently, the sequence (p_k) has a finite length, and it converges strongly toward $p_\infty = \mathrm{prox}_{\lambda f}\, x_\infty$. Using the relation $\nabla f_\lambda(x_\infty) \in \partial f(p_\infty)$, we obtain the approximate optimality property: $\partial f(p_\infty) + \partial\phi(0) \ni 0$. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Remark 4.2. It could be possible to make the parameter λ vary, but it has to be bounded away from zero (because of the assumption $\lambda \geq \frac{1}{\gamma}(\frac{h}{2}+\beta)$). Thus our approach differs from the classical approximation method which consists in approaching f by f_λ as λ goes to zero.

5. Splitting algorithms for the Lasso-type problems. In many situations, the minimization problem has an additive composite structure $\min_{\mathcal{H}}(f+g)$, with f smooth and g nonsmooth. Accelerated proximal-gradient algorithms are effective splitting methods to treat these problems. We first adapt (IPAHDD) to such a composite setting, in the case of the Lasso-type problems. Then, based on this approach, we perform some numerical experiments.

5.1. (IPAHDD) for Lasso-type problems. Take $\mathcal{H} = \mathbb{R}^n$ with the usual Euclidean structure. Suppose that the function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ to be minimized has the additive structure

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + g(x), \tag{5.1}$$

where $A \in \mathbb{R}^{m\times n}$ (with $m \leq n$), $b \in \mathbb{R}^m$ and $g \in \Gamma_0(\mathbb{R}^n)$ (set of convex, lower semicontinuous, and proper functions). Minimizing such function f occurs in a variety of fields ranging from inverse problems in signal/image processing to machine learning and statistics. Typical examples of function g include the ℓ_1 norm (Lasso), the $\ell_1 - \ell_2$ norm (group Lasso), the total variation, or the nuclear norm. In all these situations, g is nonsmooth. A direct application of (IPAHDD-nonsmooth) would require calculating (at least approximately) the proximal operator of f. It's not easy in general. To work around this difficulty, we use a change of metric. This technique was initiated by Lemaréchal and Sagastizábal in [35] to introduce efficient preconditioners into the proximal point algorithm for minimizing convex functions; for recent developments see [2], [13], [31, section 4.6]. For a symmetric and positive definite matrix $M \in \mathbb{R}^{n\times n}$, we denote by $\langle \cdot,\cdot\rangle_M = \langle M\cdot,\cdot\rangle$ the scalar product on $\mathbb{R}^n$ induced by M, and by $\|\cdot\|_M$ the associated norm. For a given $f \in \Gamma_0(\mathbb{R}^n)$, the Moreau's envelope of index $\lambda > 0$ associated with the metric induced by M is the function $f_\lambda^M : \mathcal{H} \to \mathbb{R}$ defined by, for $x \in \mathbb{R}^n$,

$$f_\lambda^M(x) = \min_{\xi\in\mathbb{R}^n} \left\{ f(\xi) + \frac{1}{2\lambda}\|x - \xi\|_M^2 \right\}. \tag{5.2}$$

Let us denote by $\mathrm{prox}_{\lambda f}^M(x)$ the unique minimizer in (5.2), which is the proximal point of x, of index λ, for the metric induced by M. The first-order optimality condition for (5.2) gives

$$\mathrm{prox}_{\lambda f}^M(x) = (M + \lambda\partial f)^{-1}(Mx). \tag{5.3}$$

When $M = I_n$ (the identity matrix), we find the classical definitions. It is easy to prove that

$$\left\|\mathrm{prox}_{\lambda f}^M(x_1) - \mathrm{prox}_{\lambda f}^M(x_2)\right\| \leq \frac{\mu_{\max}(M)}{\mu_{\min}(M)}\|x_1 - x_2\|,$$

where $\mu_{\max}(M)$ and $\mu_{\min}(M)$ are respectively the largest and the smallest eigenvalue of M. The Moreau envelope f_λ^M is a $C^{1,1}$ function whose gradient for the Euclidean structure is given by

$$\nabla f_\lambda^M(x) = \frac{1}{\lambda}M\Big(x - \mathrm{prox}_{\lambda f}^M(x)\Big). \tag{5.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

As a classical result, ∇f_λ^M is $\frac{1}{\lambda}$-Lipschitz continuous for the norm $\|\cdot\|_M$. From this, by using classical linear algebra, we easily deduce that

$$\left\|\nabla f_\lambda^M(x_1) - \nabla f_\lambda^M(x_2)\right\| \le \frac{1}{\lambda}\sqrt{\frac{\mu_{\max}(M)}{\mu_{\min}(M)}}\|x_1 - x_2\| \quad \forall x_1,\ x_2 \in \mathbb{R}^n. \tag{5.5}$$

On the other hand, one can check easily that $\operatorname{argmin}(f_\lambda^M) = \operatorname{argmin}(f)$. With the particular choice of f in (5.1), we set $M = I_n - \lambda A^T A$. If $\lambda \in [0, \frac{1}{\|A\|_2^2}[$, then M is positive definite. In this case,

$$\operatorname{prox}_{\lambda f}^M(x) = \operatorname{prox}_{\lambda g}\Big(x - \lambda A^T(Ax - b)\Big). \tag{5.6}$$

Note that formula (5.6) for the composite optimization problem (5.1) was given in [31, section 4.6, p. 190]. Using (5.4) and (5.6), we get

$$\nabla f_\lambda^M(x) = \frac{1}{\lambda} M\Big(x - \operatorname{prox}_{\lambda g}\Big(x - \lambda A^T(Ax - b)\Big)\Big). \tag{5.7}$$

Replacing f with f_λ^M in (IPAHDD), we obtain the following splitting algorithm applicable to (5.1):

(IPAHDD)-Lasso

Initialize: $x_0 \in \mathbb{R}^n$, $x_1 \in \mathbb{R}^n$, $M = I_n - \lambda A^T A,\ 0 < \lambda\|A\|_2^2 < 1$

$z_k = \frac{1}{\lambda} M\Big(x_k - \operatorname{prox}_{\lambda g}\Big(x_k - \lambda A^T(Ax_k - b)\Big)\Big)$

$y_k = \frac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \frac{\beta}{1+h\gamma}(z_k - z_{k-1}) - \frac{h}{1+h\gamma} z_k$

$x_{k+1} = x_k + h \operatorname{prox}_{\frac{h}{1+h\gamma}\phi}(y_k)$

A direct application of Theorems 2.1–2.2 to the model situation $\phi(x) = r\|x\|_2$ gives the following.

THEOREM 5.1. *Assume that the potential friction function $\phi = r\|\cdot\|_2$. Suppose that the parameters h, γ, β, λ in the algorithm* (IPAHDD)-Lasso *satisfy the relation*

$$\gamma \ge \frac{1}{\lambda}\sqrt{\frac{\mu_{\max}(M)}{\mu_{\min}(M)}}\left(\frac{h}{2} + \beta\right).$$

Let (x_k) be a sequence defined by the algorithm (IPAHDD)-Lasso. *Then,*

$$\sum_k \|x_{k+1} - x_k\| < +\infty, \text{ and therefore } \lim_k x_k := x_\infty \text{ exists.}$$

Moreover, $\sum_{k=1}^\infty \|x_{k+1} - x_k\| \le \frac{1}{r}(\frac{1}{2}\|\frac{1}{h}(x_1 - x_0)\|^2 + (f(x_1) - \inf_{\mathcal{H}} f) + \frac{\beta L}{2h}\|x_1 - x_0\|^2)$. The limit x_∞ satisfies $\|A^T(Ax_\infty - b) + \partial g(x_\infty)\| \le r$.

Suppose that the above inequality is strict. Then, (x_k) is finitely convergent.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

5.2. Some numerical experiments. Let us perform some numerical tests to compare the four algorithms IPAHDD, IPAHDD-Var, IPAHDD-N, and IPAHDD-N-Var defined in the last sections, as well as the well-known algorithms ISTA and FISTA; see [24]. We use the *performance profiles* developed by Dolan and Moré [32] as a tool for comparing the solvers. The performance profiles give for each $t \in \mathbb{R}$ the proportion $\rho_s(t)$ of test problems on which each solver s under comparison has a performance within the factor t of the best possible ratio. For more details, we refer to [32]. To compare these algorithms, we choose the number of iterations found by each solver as a performance measure. The function $\phi : \mathbb{R}^n \to \mathbb{R}$ is given by $x \mapsto \phi(x) = r\|x\|_2$ with $r = 0.1$, while the functions $f : \mathbb{R}^n \to \mathbb{R}$ are quadratic of the form $f(x) = \frac{1}{2}\|Ax - b\|^2$, $A \in \mathbb{R}^{m\times n}$ (with $m \leq n$) and $b \in \mathbb{R}^m$ are chosen randomly. The matrices A in our set of tests come from the SuiteSparse Matrix Collection.[3] We have chosen a set P of 42 different problems with matrices $A \in \mathbb{R}^{m\times n}$ size ranging from $m = 24$ to $m = 1309$ and from $n = 1309$ to $n = 1706$. The numerical experiments are carried out on an iMac with Mac OS 10.14 and a 3.2 GHz Intel Xeon W processor with 64 Go memory. All the codes are written and executed in MATLAB R2018b. We use the same initial points and the same stopping criterion, i.e., either the number of iterations exceeds 10^5 or $\|\nabla f(x_k)\| \leq r$. We observe in Figure 5.1 (left) that solvers IPAHDD, IPAHDD-Var, IPAHDD-N, and IPAHDD-N-Var are robust and solved all problems. The algorithm IPAHDD-N-Var is the most efficient. The algorithms IPAHDD and IPAHDD-Var solve 80% of the problems in the interval $[0, 1.5]$, while IPAHDD-N is robust after $t \geq 4.5$. We observe also that the algorithm FISTA solves 95% of the problems and is robust, while ISTA solves less than 60% of the problems. We conclude that, using the same initial points and under the same stopping criteria, IPAHDD-N-Var is the winner, followed by IPAHDD and IPAHDD-Var. In order to measure the effect of the introduction of the Hessian-driven viscous damping β, we also tested the four algorithms IPAHDD, IPAHDD-Var, IPAHDD-N, and IPAHDD-N-Var with $\beta = 0$ and $\beta = 0.3$. We observe that for all four methods, the introduction of the Hessian-driven viscous damping $\beta > 0$ has favorable effects not only for the convergence of the algorithm but also for the acceleration of the convergence. We also compared the algorithm IPAHDD for the LASSO problem with $g(x) = \|x\|_1$ where formula (2.27) is used to compute $\text{prox}_{\lambda g}$. Figure 5.1 (right) shows the effect of the introduction of the Hessian-driven viscous damping. Consistent with the theoretical part, we observe

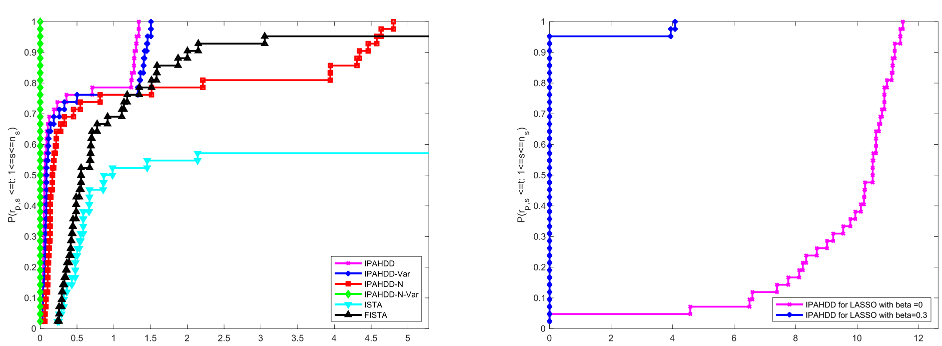

FIG. 5.1. *Performance profiles with $t_{p,s}$ the number of iterations (left). Performance profiles for IPAHDD for the LASSO problem with $t_{p,s}$ the number of iterations (right).*

[3]https://sparse.tamu.edu.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

that both dry friction coefficient $r > 0$ and Hessian-driven viscous damping coefficient $\beta > 0$ introduce some stability, robustness, and acceleration of the convergence in the numerical algorithms studied in this paper. We also have performed some numerical simulations taking into account the cpu time as a performance measure. The results are very similar to those presented for the number of iterations in Figure 5.1 and the figures have been omitted due to the limitation of the number of pages.

6. Perspectives. Let's list some of the many directions of research for the future:

(i) The algorithm (IPAHDD) works with a general smooth function f, without any convexity assumption on f. We were able to extend our study to the case of a nonsmooth convex function, by using the properties of the Moreau envelope in the convex case. For a nonconvex nonsmooth function, a natural idea would be to use the Lasry–Lions regularization for which similar regularity properties are valid.

(ii) The robustness and tolerance to perturbations and errors of the algorithm (IPAHDD) naturally suggest developing corresponding stochastic gradient methods.

(iii) We have developed our method in the case of Lasso-type problems. It would be interesting to further develop the method in order to capture a larger class of composite problems.

(iv) It would be interesting to consider the nonautonomous case where the viscous damping coefficient $\gamma(t)$ tends to zero like $\frac{\alpha}{t}$, $\alpha \geq 3$. This would allow dry friction and Hessian-driven damping to be considered together with Nesterov's fast gradient method.

(v) A natural extension would be to consider the case of monotone inclusions. The closed-loop stopping rule provided by dry friction fits well with this class of problems.

Appendix A. Proof of Theorem 2.7. The beginning of the proof is similar to that of Theorem 2.1. Taking the scalar product of (2.31) with $\frac{1}{h}(x_{k+1} - x_k)$, we obtain

$$\begin{aligned}
&\left\langle \frac{1}{h}(x_{k+1} - x_k) - \frac{1}{h}(x_k - x_{k-1}),\, \frac{1}{h}(x_{k+1} - x_k) \right\rangle + \gamma h \left\| \frac{1}{h}(x_{k+1} - x_k) \right\|^2 \\
&\quad + h \left\langle \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right), \frac{1}{h}(x_{k+1} - x_k) \right\rangle + \beta \left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1} - x_k) \right\rangle \\
&\quad + \langle \nabla f(x_k), x_{k+1} - x_k \rangle = \langle e_k, x_{k+1} - x_k \rangle .
\end{aligned} \tag{A.1}$$

Set $X_k := \frac{1}{h}(x_k - x_{k-1})$. Using convex subdifferential inequalities, we get

$$\begin{aligned}
&\frac{1}{2}\|X_{k+1}\|^2 - \frac{1}{2}\|X_k\|^2 + \gamma h \|X_{k+1}\|^2 + h\phi(X_{k+1}) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\
&\quad + \beta \left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1} - x_k) \right\rangle \leq \langle e_k, x_{k+1} - x_k \rangle .
\end{aligned} \tag{A.2}$$

According to the assumption $(\mathrm{DF})_r$ on ϕ and Lemma 1.1, for all $k \geq 1$

$$\phi(X_{k+1}) \geq r \|X_{k+1}\|. \tag{A.3}$$

Since ∇f is L-Lipschitz continuous, the classical gradient descent gives, for all $k \geq 1$,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{A.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

According to the Cauchy–Schwarz inequality, and using again that ∇f is L-Lipschitz continuous,

$$\left|\left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1}-x_k)\right\rangle\right| \le hL\|X_k\|\|X_{k+1}\| \le \frac{hL}{2}(\|X_k\|^2 + \|X_{k+1}\|^2). \tag{A.5}$$

Combining (A.3)–(A.4)–(A.5) with (A.2), and using the Cauchy–Schwarz inequality, we obtain

$$\frac{1}{2}\left\|\frac{1}{h}(x_{k+1}-x_k)\right\|^2 - \frac{1}{2}\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2 + \frac{\gamma}{h}\|x_{k+1}-x_k\|^2 + r\|x_{k+1}-x_k\|$$
$$+ f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1}-x_k\|^2 \le \frac{\beta L}{2h}(\|x_k-x_{k-1}\|^2 + \|x_{k+1}-x_k\|^2)$$
$$+ \|e_k\|\|x_{k+1}-x_k\|.$$

In terms of $E_k := \frac{1}{2}\|\frac{1}{h}(x_k-x_{k-1})\|^2 + (f(x_k) - \inf f)$, this is equivalent to

$$E_{k+1} - E_k + \left(\frac{\gamma}{h} - \frac{L}{2} - \frac{\beta L}{2h}\right)\|x_{k+1}-x_k\|^2 + (r - \|e_k\|)\|x_{k+1}-x_k\| \le \frac{\beta L}{2h}\|x_k-x_{k-1}\|^2. \tag{A.6}$$

According to the assumption $\gamma \ge L\left(\frac{h}{2}+\beta\right)$, we have $\frac{\gamma}{h} - \frac{L}{2} - \frac{\beta L}{2h} \ge \frac{\beta L}{2h}$. Therefore,

$$E_{k+1} - E_k + (r - \|e_k\|)\|x_{k+1}-x_k\| + \frac{\beta L}{2h}\|x_{k+1}-x_k\|^2 \le \frac{\beta L}{2h}\|x_k-x_{k-1}\|^2. \tag{A.7}$$

Set $\tilde{E}_k := E_k + \frac{\beta L}{2h}\|x_k-x_{k-1}\|^2$. We have

$$\tilde{E}_{k+1} - \tilde{E}_k + (r - \|e_k\|)\|x_{k+1}-x_k\| \le 0. \tag{A.8}$$

Adding the above inequalities, and according to $E_k \ge 0$, and $e_k \to 0$, we deduce from (2.11) that $\sum_{k=1}^{\infty}\|x_{k+1}-x_k\| < +\infty$. So, the sequence (x_k) has a finite length, which implies that the strong limit of (x_k) exists. Set $x_\infty := \lim x_k$. Therefore, $\lim_k \|x_{k+1}-x_k\| = 0$, $\lim_k \frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1}) = 0$, and $\lim_k \nabla f(x_k) = \nabla f(x_\infty)$. To pass to the limit on (2.31), rewrite it as follows: $A_k \in \partial\phi(\frac{1}{h}(x_{k+1}-x_k))$ with

$$A_k = -\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1}) - \frac{\gamma}{h}(x_{k+1}-x_k) - \frac{\beta}{h}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \nabla f(x_k) + e_k.$$

According to the above convergence results and the closedness of the graph of $\partial\phi$, we deduce that

$$-\nabla f(x_\infty) \in \partial\phi(0).$$

The proof of (iii) and (iv) follows the lines of the proof of Theorem 2.2. Estimation (2.17) becomes

$$\frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_{k+1}-x_k)\right\|^2 - \frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2 + \frac{1}{h}(\gamma-\beta L)\|x_{k+1}-x_k\|^2 + \varepsilon\|x_{k+1}-x_k\| \le \|e_k\|\|x_{k+1}-x_k\|. \tag{A.9}$$

Since $\|e_k\| \to 0$, we obtain that for k sufficiently large

$$\frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_{k+1}-x_k)\right\|^2 - \frac{1}{2}(1+\beta hL)\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2$$
$$+ \frac{1}{h}(\gamma-\beta L)\|x_{k+1}-x_k\|^2 + \frac{\varepsilon}{2}\|x_{k+1}-x_k\| \le 0.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

According to $\gamma - \beta L > 0$, we easily deduce the geometric convergence of the sequence (x_k).

To prove the finite convergence, we return to the definition of the algorithm (IPAHDD-pert):

$$\frac{1}{h}(x_{k+1} - x_k) = \mathrm{prox}_{\lambda\phi}(\xi_k), \text{ where } \lambda = \frac{h}{1+h\gamma}, \text{ and}$$

$$\xi_k := \tfrac{1}{h(1+h\gamma)}(x_k - x_{k-1}) - \tfrac{\beta}{1+h\gamma}(\nabla f(x_k) - \nabla f(x_{k-1})) - \tfrac{h}{1+h\gamma}\nabla f(x_k) + \tfrac{h}{1+h\gamma}e_k.$$

According to Lemma 1.2, the finite convergence will result from proving the inequality $\frac{1}{\lambda}\|\xi_k\| \le r$. Since $\frac{1}{\lambda}\|\xi_k\| \to \|\nabla f(x_\infty)\|$, this will be satisfied for k large enough if $\|\nabla f(x_\infty)\| < r$.

Appendix B. Proof of Theorem 3.1. (i) By taking the scalar product of (3.2) with $\frac{1}{h}(x_{k+1} - x_k)$, we obtain

$$\begin{aligned} &\left\langle \frac{1}{h}(x_{k+1} - x_k) - \frac{1}{h}(x_k - x_{k-1}), \frac{1}{h}(x_{k+1} - x_k)\right\rangle + \gamma h \left\|\frac{1}{h}(x_{k+1} - x_k)\right\|^2 \\ &\quad + \langle \nabla f(y_k), x_{k+1} - x_k\rangle + h\left\langle \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right), \frac{1}{h}(x_{k+1} - x_k)\right\rangle \\ &\quad + \beta\left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1} - x_k)\right\rangle = 0. \end{aligned} \tag{B.1}$$

Set $X_k := \frac{1}{h}(x_k - x_{k-1})$. According to the assumption $(\mathrm{DF})_r$ on ϕ and Lemma 1.1, for all $k \ge 1$

$$\langle \partial\phi(X_{k+1}), X_{k+1}\rangle \ge \phi(X_{k+1}) \ge r\|X_{k+1}\|.$$

According to the Cauchy–Schwarz inequality, and using that ∇f is L-Lipschitz continuous,

$$\left|\left\langle \nabla f(x_k) - \nabla f(x_{k-1}), \frac{1}{h}(x_{k+1} - x_k)\right\rangle\right| \le hL\|X_k\|\|X_{k+1}\| \le \frac{hL}{2}(\|X_k\|^2 + \|X_{k+1}\|^2).$$

Combining (B.1) with the two above inequalities, we obtain

$$\begin{aligned} &\langle X_{k+1} - X_k, X_{k+1}\rangle + \gamma h\|X_{k+1}\|^2 + hr\|X_{k+1}\| + \langle \nabla f(y_k), x_{k+1} - x_k\rangle \\ &\quad \le \frac{\beta hL}{2}(\|X_k\|^2 + \|X_{k+1}\|^2). \end{aligned}$$

Using successively the gradient descent lemma for f, the L-Lipschitz continuity of ∇f, and the equality $y_k = x_k + \frac{1}{1+h\gamma}(x_k - x_{k-1})$ (by definition of (IPAHDD-N)), we get

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\le f(x_k) + \langle \nabla f(y_k), x_{k+1} - x_k\rangle + \frac{L}{1+h\gamma}\|x_k - x_{k-1}\|\|x_{k+1} - x_k\| + \frac{L}{2}\|x_{k+1} - x_k\|^2. \end{aligned}$$

Combining the above inequalities, we obtain

$$\begin{aligned} &\langle X_{k+1} - X_k, X_{k+1}\rangle + \gamma h\|X_{k+1}\|^2 + hr\|X_{k+1}\| + f(x_{k+1}) - f(x_k) \\ &\quad - \frac{Lh^2}{1+h\gamma}\|X_k\|\|X_{k+1}\| - \frac{Lh^2}{2}\|X_{k+1}\|^2 \le \frac{\beta hL}{2}(\|X_k\|^2 + \|X_{k+1}\|^2). \end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Therefore,

$$
\begin{aligned}
&\left(1+h\gamma-\frac{Lh^2}{2}-\frac{\beta hL}{2}\right)\|X_{k+1}\|^2-\left(1+\frac{Lh^2}{1+h\gamma}\right)\|X_k\|\|X_{k+1}\|-\frac{\beta hL}{2}\|X_k\|^2 \\
&\quad + hr\|X_{k+1}\|+f(x_{k+1})-f(x_k)\le 0.
\end{aligned}
\tag{B.2}
$$

According to $\gamma \geq \frac{3L}{2}(h+\beta)$, and $Lh^2\leq 1$, we get $1+h\gamma-\frac{Lh^2}{2}-\frac{\beta hL}{2}\geq 0$. From (B.2) we infer

$$
\begin{aligned}
&\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{\beta hL}{2}\right)(\|X_{k+1}\|^2-\|X_k\|^2)+\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{\beta hL}{2}\right)\|X_{k+1}\|^2 \\
&\quad +\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{3\beta hL}{2}\right)\|X_k\|^2-\left(1+\frac{Lh^2}{1+h\gamma}\right)\|X_k\|\|X_{k+1}\| \\
&\quad + hr\|X_{k+1}\|+f(x_{k+1})-f(x_k)\le 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{\beta hL}{2}\right)(\|X_{k+1}\|^2-\|X_k\|^2)+\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{3\beta hL}{2}\right)\|X_{k+1}\|^2 \\
&\quad +\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{3\beta hL}{2}\right)\|X_k\|^2-\left(1+\frac{Lh^2}{1+h\gamma}\right)\|X_k\|\|X_{k+1}\| \\
&\quad + hr\|X_{k+1}\|+f(x_{k+1})-f(x_k)\le 0.
\end{aligned}
$$

Elementary algebra (sign of a polynomial of the second degree) gives that the inequality

$$
\begin{aligned}
&\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{3\beta hL}{2}\right)\|X_{k+1}\|^2-\left(1+\frac{Lh^2}{1+h\gamma}\right)\|X_k\|\|X_{k+1}\| \\
&\quad +\frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}-\frac{3\beta hL}{2}\right)\|X_k\|^2\ge 0
\end{aligned}
$$

is satisfied under the condition $\Delta = (1+\frac{Lh^2}{1+h\gamma})^2-(1+h\gamma-\frac{Lh^2}{2}-\frac{3\beta hL}{2})^2\leq 0$. This is equivalent to $\frac{\gamma}{L}\geq \frac{h}{1+h\gamma}+\frac{h}{2}+\frac{3\beta}{2}$. Since $\frac{1}{2}+\frac{1}{1+h\gamma}\leq\frac{3}{2}$, we end up with the condition $\frac{\gamma}{L}\geq\frac{3}{2}(h+\beta)$, which is satisfied by assumption. To summarize the results, in terms of

$$
E_k := \frac{1}{2}\left(1+h\gamma-\frac{Lh^2}{2}\right)\left\|\frac{1}{h}(x_k-x_{k-1})\right\|^2+\left(f(x_k)-\inf_{\mathcal{H}} f\right),
$$

we have obtained $E_{k+1}-E_k+r\|x_{k+1}-x_k\|\leq 0$. According to the nonnegativity of E_k, and $r>0$, we deduce that $\sum_{k=1}^{\infty}\|x_{k+1}-x_k\|\leq\frac{1}{r}E_1<+\infty$. Therefore, the sequence (x_k) has a finite length, which implies that the strong limit of (x_k) exists. Set $x_\infty := \lim_k x_k$, which ends item (i).

(ii) From $\sum_k\|x_{k+1}-x_k\|<+\infty$, we get immediately $\lim_k\|x_{k+1}-x_k\|=0$. This in turn implies $\lim_k\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})=\lim_k\frac{1}{h^2}((x_{k+1}-x_k)-(x_k-x_{k-1}))=0$. Moreover, since ∇f is continuous and (x_k) converges strongly to x_∞, we have $\nabla f(x_k)\to\nabla f(x_\infty)$. According to the L-Lipschitz continuity of ∇f, and $y_k-x_k=\frac{1}{1+h\gamma}(x_k-x_{k-1})$, we have

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\|\nabla f(y_k) - \nabla f(x_k)\| \leq L\|y_k - x_k\| \leq \frac{L}{1+h\gamma}\|x_k - x_{k-1}\|.$$

Therefore, $\lim_k \nabla f(y_k) = \nabla f(x_\infty)$. To pass to the limit in (3.1), rewrite it as follows:

$$-\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) - \frac{\gamma}{h}(x_{k+1} - x_k) - \nabla f(y_k) \in \partial\phi\left(\frac{1}{h}(x_{k+1} - x_k)\right).$$

According to the above convergence results and the closedness of the graph of $\partial\phi$ in $\mathcal{H} \times \mathcal{H}$, we deduce that $-\nabla f(x_\infty) \in \partial\phi(0)$, which gives item (ii). Items (iii) and (iv) are obtained by a similar argument as in Theorem 2.2. By definition of (IPAHDD-N), we have $\|y_k - x_k\| \leq \|x_k - x_{k-1}\|$, and hence $y_k - x_k$ converges strongly to zero. From this, we immediately deduce that the sequence (y_k) has the same limit as (x_k) and satisfies similar convergence properties as (x_k).

Acknowledgment. The authors would like to thank the three anonymous reviewers for their careful reading and their relevant suggestions and comments.

REFERENCES

[1] S. Adly, *A Variational Approach to Nonsmooth Dynamics: Applications in Unilateral Mechanics and Electronics*, Springer Briefs in Math., Springer, New York, 2017.

[2] S. Adly and H. Attouch, *Finite Convergence of Proximal-Gradient Inertial Algorithms with Dry Friction Damping*, https://hal.archives-ouvertes.fr/hal-02388038, 2019.

[3] S. Adly and H. Attouch, *Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping*, J. Convex Anal., 28 (2021).

[4] S. Adly, H. Attouch, and A. Cabot, *Finite time stabilization of nonlinear oscillators subject to dry friction*, in Nonsmooth Mechanics and Analysis, Adv. Mech. Math. 12, Springer, New York, 2006, pp. 289–304.

[5] F. Álvarez, *On the minimizing property of a second-order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[6] F. Álvarez, H. Attouch, J. Bolte, and P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian-driven damping*, J. Math. Pures Appl., 81 (2002), pp. 747–779.

[7] H. Amann and J. I. Díaz, *A note on the dynamics of an oscillator in the presence of strong friction*, Nonlinear Anal., 55 (2003), pp. 209–216.

[8] V. Apidopoulos, J.-F. Aujol, and Ch. Dossal, *Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule*, Math. Program., 180 (2018).

[9] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, 2nd ed., MOS/SIAM Ser. Optim. 17, SIAM, Philadelphia, 2014.

[10] H. Attouch and A. Cabot, *Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity*, J. Differential Equations, 263 (2017), pp. 5412–5458.

[11] H. Attouch and A. Cabot, *Convergence rates of inertial forward-backward algorithms*, SIAM J. Optim., 28 (2018), pp. 849–874.

[12] H. Attouch, A. Cabot, Z. Chbani and H. Riahi, *Rate of convergence of inertial gradient dynamics with time-dependent viscous damping coefficient*, Evol. Equ. Control Theory, 7 (2018), pp. 353–371.

[13] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi, *First-Order Optimization Algorithms via Inertial Systems with Hessian Driven Damping*, https://hal.archives-ouvertes.fr/hal-02193846, 2019.

[14] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program. Ser. B, 168 (2018), pp. 123–175.

[15] H. Attouch, Z. Chbani, and H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM J. Optim., 29 (2019), pp. 2227–2256.

[16] H. Attouch, Z. Chbani, and H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, ESAIM Control Optim. Calc. Var., 25 (2019).

[17] H. Attouch, X. Goudou, and P. Redont, *The heavy ball with friction method. The continuous dynamical system, global exploration of the local minima of a real-valued function*

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/25/20 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

by asymptotical analysis of a dissipative dynamical system, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[18] H. Attouch and J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$*, SIAM J. Optim., 26 (2016), pp. 1824–1834.

[19] H. Attouch, J. Peypouquet, and P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261 (2016), pp. 5734–5783.

[20] J.-F. Aujol and Ch. Dossal, *Stability of over-relaxations for the Forward-Backward algorithm: Application to FISTA*, SIAM J. Optim., 25 (2015), pp. 2408–2433.

[21] J.-F. Aujol and Ch. Dossal, *Optimal Rate of Convergence of an ODE Associated to the Fast Gradient Descent Schemes for $b > 0$*, https://hal.inria.fr/hal-01547251v2, 2017.

[22] B. Baji and A. Cabot, *An inertial proximal algorithm with dry friction: Finite convergence results*, Set-Valued Anal., 9 (2006), pp. 1–23.

[23] H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Math., Springer, New York, 2011.

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

[25] R. I. Bot, E. R. Csetnek, and S. C. László, *A second order dynamical approach with variable damping to nonconvex smooth minimization*, Appl. Anal., 99 (2020), pp. 361–378.

[26] R. I. Bot, E. R. Csetnek, and S. C. László, *Tikhonov Regularization of a Second Order Dynamical System with Hessian Driven Damping*, arXiv:1911.12845v1[math.OC], 2019.

[27] R. I. Bot and E. R. Csetnek, *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM J. Control Optim., 54 (2016), pp. 1423–1443.

[28] R. I. Bot and E. R. Csetnek, *A second order dynamical system with Hessian-driven damping and penalty term associated to variational inequalities*, Optimization, 68 (2019), pp. 1265–1277.

[29] H. Brézis, *Opérateurs maximaux monotones dans les espaces de Hilbert et équations d'évolution*, Lecture Notes 5, North-Holland, Amsterdam, 1972.

[30] A. Chambolle and Ch. Dossal, *On the convergence of the iterates of the fast iterative shrinkage thresholding algorithm*, J. Optim. Theory Appl., 166 (2015), pp. 968–982.

[31] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numer., 25 (2016), pp. 161–319.

[32] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.

[33] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, *Global convergence of the heavy-ball method for convex optimization*, in Proceedings of the European Control Conference, 2015, pp. 310–315.

[34] S. C. László, *Forward-Backward Algorithms with Different Inertial Terms for the Minimization of the Sum of Two Non-Convex Functions*, arXiv:2002.07154v2[math.FA], 2020.

[35] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau-Yosida regularization: Theoretical preliminaries.* SIAM J. Optim., 7 (1997), pp. 367–385.

[36] J. Liang, J. Fadili, and G. Peyré, *Local linear convergence of forward-backward under partial smoothness*, in Advances in Neural Information Processing Systems, 2014, pp. 1970–1978.

[37] R. May, *Asymptotic for a second-order evolution equation with convex potential and vanishing damping term*, Turkish J. Math., 41 (2017), pp. 681–685.

[38] M. Marques Alves, R. Monteiro, and B. F. Svaiter, *Regularized HPE-type methods for solving monotone inclusions with improved pointwise iteration complexity bounds*, SIAM. J. Optim., 26 (2016), pp. 2730–2743.

[39] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Math. Dokl., 27 (1983), pp. 372–376.

[40] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic Publishers, Boston, MA, 2004.

[41] B. T. Polyak, *Some methods of speeding up the convergence of iterative methods*, Z. Vylist Math. Fiz., 4 (1964), pp. 1–17.

[42] B. T. Polyak, *Introduction to Optimization.* Optimization Software, New York, 1987.

[43] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, *Understanding the Acceleration Phenomenon via High-Resolution Differential Equations*, arXiv:submit/2440124[cs.LG], 2018.

[44] J. W. Siegel, *Accelerated First-Order Methods: Differential Equations and Lyapunov Functions*, arXiv:1903.05671v1[math.OC], 2019.

[45] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.