First order inertial optimization algorithms with threshold effects associated with dry friction

Samir Adly, Hedy Attouch, Manh Hung Le

First order inertial optimization algorithms with threshold effects associated with dry friction

Samir Adly* Hedy Attouch† Manh Hung Le ‡

July 12, 2021

ABSTRACT. In a Hilbert space setting, we consider a new first order optimization algorithm which is obtained by temporal discretization of a damped inertial dynamic involving dry friction. The function f to be minimized is assumed to be differentiable (not necessarily convex). The dry friction potential function ϕ, which has a sharp minimum at the origin, enters the algorithm via its proximal mapping, which acts as a soft thresholding operator on the sum of the velocity and the gradient terms. After a finite number of steps, the structure of the algorithm changes, losing its inertial character to become the steepest descent method. The geometric damping driven by the Hessian of f makes it possible to control and attenuate the oscillations. The algorithm generates convergent sequences when f is convex, and in the nonconvex case when f satisfies the Kurdyka-Lojasiewicz property. As a remarkable property, the convergence results tolerate the presence of errors, under the sole assumption of their asymptotic convergence towards zero. The study is then extended to the case of a nonsmooth convex function f, in which case the algorithm involves the proximal operators of f and ϕ separately. Then, applications are given to the Lasso problem and nonsmooth d.c. programming.

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

Key words and phrases: proximal-gradient algorithms; inertial methods; dry friction; Hessian-driven damping; soft thresholding; Kurdyka-Lojasiewicz property; Lasso problem; d.c. optimization; errors.

1 Introduction

Throughout the paper, $\mathcal{H}$ is a real Hilbert space equipped with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. The objective function $f : \mathcal{H} \to \mathbb{R}$ is assumed to be differentiable with Lipschitz continuous gradient. Unless otherwise specified, f is not assumed to be convex. When we consider the continuous dynamic on which the algorithms are based, and where the Hessian is involved, more regularity is needed for f which is then assumed to be C^2. Weakening these assumptions by removing the smoothness of the objective function f will be examined at the end of the paper. We will analyze the convergence properties of several algorithms that can be obtained by temporal discretization of the differential inclusion

$$\boxed{\ddot{x}(t) + \gamma \dot{x}(t) + \partial \phi \Big(\dot{x}(t) + \beta \nabla f(x(t)) \Big) + \beta \nabla^2 f(x(t)) \dot{x}(t) + \nabla f(x(t)) \ni 0,} \tag{1.1}$$

*Laboratoire XLIM, Université de Limoges, 123, avenue Albert Thomas, 87060 Limoges, France. E-mail: samir.adly@unilim.fr

†IMAG, Université Montpellier, CNRS, Place Eugène Bataillon, 34095 Montpellier CEDEX 5, France. E-mail: hedy.attouch@umontpellier.fr, Supported by COST Action: CA16228

‡Laboratoire XLIM, Université de Limoges, 123, avenue Albert Thomas, 87060 Limoges, France. E-mail: manh-hung.le@etu.unilim.fr

where $\gamma > 0$ and $\beta > 0$ are respectively the viscous damping and Hessian damping coefficients, and ϕ is a dry friction potential function with a sharp minimum at the origin. This type of autonomous system, with a damping which acts as a closed loop control of the sum of the velocity and gradient terms, was recently introduced by Attouch, Bot, and Csetnek in [9]. It falls within the general framework of the use of inertial dynamics in optimization to accelerate algorithms, as mechanical intuition naturally suggests. An abundant literature has been devoted to the link between damped inertial dynamics and corresponding optimization algorithms obtained by temporal discretization, see e.g. [11, 14, 26, 27, 42, 44, 48] for recent developments on the subject. The term $\gamma\dot{x}(t)$ in (1.1) models the viscous damping with a fixed positive coefficient $\gamma > 0$. The case where the viscous damping coefficient is time dependent and vanishes asymptotically ($\gamma(t) \to 0$ as $t \to +\infty$) is of particular interest due to its connection with the Nesterov acceleration method, see [48]. In fact, our article deals with the case of a fixed viscous damping coefficient. This framework is well adapted to dry friction, and, as we will see, allows minimal assumptions in the presence of error terms in the algorithms.

Dry friction Following [1–3], we say that the potential function ϕ satisfies the dry friction property $(\mathrm{DF})_r$, $r > 0$, if the following properties are satisfied:

$$(\mathrm{DF})_r \quad \begin{cases} \phi : \mathcal{H} \to \mathbb{R}_+ \text{ is convex continuous,} \\ \min_{\xi \in \mathcal{H}} \phi(\xi) = \phi(0) = 0, \\ \phi(\xi) \geq r\|\xi\| \ \ \forall \xi \in \mathcal{H}. \end{cases}$$

The function $\phi(x) = r\|x\|$, $r > 0$ is a model example of potential which satisfies the dry friction property. In what follows, the friction potential function ϕ is assumed to satisfy the dry friction property. An important property associated with dry friction is stated in the lemma below (see [1–3] for further details).

Lemma 1.1 *Suppose that $\phi : \mathcal{H} \to \mathbb{R}_+$ satisfies $(\mathrm{DF})_r$. Then one has $\overline{\mathbb{B}}(0, r) \subset \partial\phi(0)$, and therefore*

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

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 $\lambda > 0$

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

Lemma 1.1 establishes a thresholding property for the proximal operator associated with a dry friction potential. It will play a key role in showing that after a finite number of steps our algorithm will arrive at the regime of the steepest descent method. As a specific property of (1.1), the dry friction term $\partial\phi\Big(\dot{x}(t) + \beta\nabla f(x(t))\Big)$ involves both the velocity vector and the gradient of f. This makes this dynamic different from that studied in [1], where the term of dry friction concerns only the velocity vector. A major advantage of considering the dry friction term in this new form compared to that in [1] is that the iterates generated by our algorithm will converge towards a critical point of f (a minimizer in the case where f is convex). By contrast, for each sequence (x_k) generated by the algorithms in [1], there is only convergence of (x_k) towards an "approximate" critical point x_∞ of f, that is, $-\nabla f(x_\infty) \in \partial\phi(0)$.

Dry friction is an important subject in mechanics. It produces stabilization of mechanical systems in finite-time. This contrasts with the viscous damping that can asymptotically produce many small oscillations. The use of dry friction in optimization is a relatively new topic. First results concerning the property of finite convergence under the action of dry friction were obtained by Adly, Attouch, and Cabot [4]. Corresponding results for Partial Differential Equations have been obtained by Amann and Diaz in [6].

Hessian driven damping The combination of viscous friction with dry friction and Hessian driven damping has been considered by Adly and Attouch in [1–3]. Even if the dynamic (1.1) requires that the potential f is twice differentiable, the associated algorithm is a first-order one. In fact, since the term $\nabla^2 f(x(t))\dot{x}(t)$ is the time derivative of $\nabla f(x(t))$, we obtain that its temporal discretization contains only the gradients of f at two consecutive steps, and is therefore relevant to first-order algorithms. The Hessian driven damping has a natural connection with the strong damping property in mechanics and physics, see [34]. It helps to control and attenuate the oscillation effects that occur naturally with inertial systems. The first results involving the Hessian-driven damping concerned the dynamic with fixed viscous damping

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

see [5]. The terminology (DIN) refers to the interpretation of this system as a (regularized) Dynamic Inertial Newton method. To accelerate this system, in line with the dynamic approach to the Nesterov accelerated gradient method [48], the following dynamic with asymptotic vanishing viscous damping coefficient

$$\text{(DIN - AVD)}_{\gamma,\beta} \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0$$

was studied in [13, 16, 28, 30, 37, 39, 46]. The above system preserves the convergence properties of the Nesterov accelerated gradient method. Moreover, it provides fast convergence to zero of the gradients, and reduces the oscillatory aspects. Similar properties for the first order algorithms obtained by temporal discretization were obtained by Attouch, Chbani, Fadili and Riahi [13], and Shi, Du, Jordan, and Su [46].

Presentation of the results Our goal in this paper is to focus on various temporal discretizations of (1.1) and their links with numerical optimization. Our main results concern the convergence properties of the inertial proximal-gradient algorithm with Hessian-damping and dry friction

$$\text{(IPAHDD-C1)} \qquad \begin{cases} y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1}), \\ x_{k+1} = x_k - \beta h\nabla f(x_k) + h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\phi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)\right), \end{cases}$$

which comes from the temporal discretization with step size $h > 0$ of (1.1). (IPAHDD) is the terminology introduced by Adly and Attouch [1] for this type of algorithm, which is a shorthand for Inertial Proximal Algorithm with Hessian Damping and Dry friction. The suffix C refers to the Composite form in which the dry friction acts in (1.1). Under suitable conditions on the damping parameters γ, β and the step size h, we will show that any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1) converges weakly to a minimizer of f when f is convex, and to a critical point of f when f is a nonconvex function which satisfies the Kurdyka-Lojasiewicz property. Moreover, the sequence $(x_k)_k$ follows the steepest descent method after a finite number of steps, and the summability property is satisfied $\sum \|\nabla f(x_k)\|^2 < +\infty$. The convergence results tolerate the presence of errors, under the sole assumption of their asymptotic convergence towards zero, which makes the algorithm attractive to deal with stochastic/noisy data. When f is strongly convex, (IPAHDD-C1) achieves exponential convergence. We show that various discretizations of the dynamic (1.1) 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 finally consider corresponding splitting algorithms for composite minimization, including the case of nonsmooth nonconvex d.c. programming, and Lasso problems.

Contents In section 2, we establish the general convergence properties of the inertial proximal-gradient (IPAHDD-C1). In section 3 and 4, we successively examine the case of a general convex function f, then the case of a nonconvex function f which satisfies the Kurdyka-Lojasiewicz property. In section 5, we

show the robustness of the algorithm (IPAHDD-C1) with respect to perturbations, errors. In section 6, we examine two variants of the algorithm which has a structure similar to that of the accelerated gradient method of Nesterov. In section 7, based on the variational properties of Moreau's envelope, we extend our results to the case where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semicontinuous and proper function, and then we examine the case of nonsmooth d.c. problems. In section 8, we extend our analysis to the case of additive composite optimization problems of Lasso type, and obtain a corresponding splitting algorithm. Section 9 is devoted to numerical experiments. We complete the paper with some concluding remarks and perspectives.

2 Convergence properties of the (IPAHDD-C1) algorithm

Given a constant step size $h > 0$, we consider the following temporal discretization of (1.1), which is implicit with respect to the nonsmooth operator $\partial\phi$, and explicit with respect to the smooth operator ∇f:

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

Set $y_k := \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$, $k \geq 1$. Let us reformulate (2.1) with the help of y_k. We obtain

$$y_{k+1} + \frac{h}{1+\gamma h}\partial\phi(y_{k+1}) \ni \frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h}\nabla f(x_k).$$

Equivalently,

$$y_{k+1} = \operatorname{prox}_{\frac{h}{1+\gamma h}\phi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h}\nabla f(x_k)\right), \tag{2.2}$$

which gives $x_{k+1} = x_k - \beta h\nabla f(x_k) + h\operatorname{prox}_{\frac{h}{1+\gamma h}\phi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)\right)$.
Therefore, we obtain the following algorithm

(IPAHDD-C1)

Initialize : $x_0 \in \mathcal{H}$, $x_1 \in \mathcal{H}$.
$y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$.
$x_{k+1} = x_k - \beta h\nabla f(x_k) + h\operatorname{prox}_{\frac{h}{1+\gamma h}\phi}\left(\frac{1}{1+\gamma h}y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)\right)$.

2.1 Lyapunov analysis

We can now state our main result concerning the algorithm (IPAHDD-C1).

Theorem 2.1 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function such that $\inf_{\mathcal{H}} f > -\infty$, and whose gradient is L-Lipschitz continuous. Assume that the friction potential $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$hL \leq \frac{2\gamma}{\gamma\beta + 1}. \tag{2.3}$$

Let $(x_k)_k$ be a sequence generated by (IPAHDD-C1)*. Then, the following properties are satisfied:*

(i) $\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)=0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2<+\infty$ *and* $\sum_{k=1}^{+\infty}\|x_{k+1}-x_k\|^2<+\infty$.

Proof. Multiplying (2.1) by h and rewriting it using y_k, we obtain for $k\geq 1$

$$y_{k+1}-y_k+\gamma(x_{k+1}-x_k)+h\partial\phi(y_{k+1})+h\nabla f(x_k)\ni 0. \tag{2.4}$$

Taking the scalar product of (2.4) with y_{k+1}, we obtain

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

Equivalently

$$\begin{aligned}\underbrace{\|y_{k+1}\|^2-\langle y_k,y_{k+1}\rangle}_{A}+\underbrace{\frac{\gamma}{h}\|x_{k+1}-x_k\|^2+(\gamma\beta+1)\langle x_{k+1}-x_k,\nabla f(x_k)\rangle}_{B}\\ +h\langle\partial\phi(y_{k+1}),y_{k+1}\rangle+\beta h\|\nabla f(x_k)\|^2=0.\end{aligned} \tag{2.5}$$

1) **Estimate** $h\langle\partial\phi(y_{k+1}),y_{k+1}\rangle$. Using the convexity of ϕ and $\phi(0)=0=\min_{\mathcal{H}}\phi$, we have

$$h\langle\partial\phi(y_{k+1}),y_{k+1}\rangle\geq h\phi(y_{k+1}). \tag{2.6}$$

2) **Estimate** A. We have

$$A\geq\|y_{k+1}\|^2-\|y_k\|\|y_{k+1}\|\geq\|y_{k+1}\|^2-\frac{1}{2}\left(\|y_{k+1}\|^2+\|y_k\|^2\right)=\frac{1}{2}\|y_{k+1}\|^2-\frac{1}{2}\|y_k\|^2. \tag{2.7}$$

3) **Estimate** B. From the gradient descent lemma and the L-Lipschitz continuity of ∇f, we get

$$\begin{aligned}B&\geq\frac{\gamma}{h}\|x_{k+1}-x_k\|^2+(\gamma\beta+1)(f(x_{k+1})-f(x_k)-\frac{L}{2}\|x_{k+1}-x_k\|^2)\\ &\geq(\gamma\beta+1)(f(x_{k+1})-f(x_k))+\left(\frac{\gamma}{h}-\frac{L}{2}(\gamma\beta+1)\right)\|x_{k+1}-x_k\|^2\\ &\geq(\gamma\beta+1)(f(x_{k+1})-f(x_k)),\end{aligned} \tag{2.8}$$

where the last inequality follows from the assumption (2.3) on the parameters, which gives equivalently $\frac{\gamma}{h}-\frac{L}{2}(\gamma\beta+1)\geq 0$. By combining (2.5), (2.6), (2.7) and (2.8), we obtain

$$\frac{1}{2}\|y_{k+1}\|^2-\frac{1}{2}\|y_k\|^2+(\gamma\beta+1)(f(x_{k+1})-f(x_k))+h\phi(y_{k+1})+\beta h\|\nabla f(x_k)\|^2\leq 0. \tag{2.9}$$

Equivalently

$$E_{k+1}-E_k+h\phi(y_{k+1})+\beta h\|\nabla f(x_k)\|^2\leq 0, \tag{2.10}$$

where

$$E_k:=\frac{1}{2}\|y_k\|^2+(\gamma\beta+1)\left(f(x_k)-\inf_{x\in H}f(x)\right).$$

By summing the inequalities (2.10) from $k=1$ to N, and using that $E_k\geq 0$, we obtain

$$h\sum_{k=1}^{N}\phi\Big(\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)\Big)+\beta h\sum_{k=1}^{N}\|\nabla f(x_k)\|^2\leq E_1-E_{N+1}\leq E_1.$$

Letting $N \to +\infty$, and since h, β are supposed to be positive, we obtain

$$\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty \text{ and } \sum_{k=1}^{+\infty} \phi\Big(\frac{1}{h}(x_{k+1}-x_k) + \beta\nabla f(x_k)\Big) < +\infty. \tag{2.11}$$

Since ϕ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$, we deduce that

$$\sum_{k=1}^{+\infty} \|\frac{1}{h}(x_{k+1}-x_k) + \beta\nabla f(x_k)\| < +\infty, \text{ that is } \sum_{k=1}^{+\infty} \|y_k\| < +\infty. \tag{2.12}$$

Therefore, $\lim_k y_k = 0$, which implies $\|y_k\|^2 \leq \|y_k\|$ for k large enough, and hence $\sum_{k=1}^{+\infty} \|y_k\|^2 < +\infty$. This property, combined with $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$ immediately gives

$$\sum_{k=1}^{+\infty} \|x_{k+1}-x_k\|^2 < +\infty.$$

Let us now prove that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent method. The proof relies on Lemma 1.1. Recall that, according to (2.2), we have the following equivalent formulation of the algorithm (IPAHDD-C1)

$$y_{k+1} = \mathrm{prox}_{\frac{h}{1+\gamma h}\phi}(z_k),$$

where

$$z_k = \frac{1}{1+\gamma h} y_k + \frac{(\gamma\beta - 1)h}{1+\gamma h}\nabla f(x_k).$$

According to (2.11), and since the general term of a convergent series necessarily goes to zero,

$$\lim_k \nabla f(x_k) = \lim_k y_k = 0.$$

According to the definition of z_k, we get $\lim_k z_k = 0$. Therefore, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\|z_k\| \leq \frac{hr}{1+\gamma h}.$$

According to Lemma 1.1, this implies that $y_{k+1} = \mathrm{prox}_{\frac{h}{1+\gamma h}\phi}(z_k) = 0$ for all $k \geq k_0$. Equivalently, $\frac{1}{h}(x_{k+1}-x_k) + \beta\nabla f(x_k) = 0$, which means that after a finite number of steps, the sequence follows the steepest descent algorithm. This completes the proof. ■

Remark 2.1 When $\mathcal{H}$ is a finite dimensional Hilbert space, let us give another proof of the fact that $y_k = 0$ after a finite number of steps, *i.e.* $(x_k)_k$ follows the steepest descent method. We argue by contradiction, which leads to supposing that there exists a subsequence $(y_{k_l})_l$ such that $\|y_{k_l+1}\| > 0$ for all $l \in \mathbb{N}$. From (2.4), we have

$$-\frac{1}{h}(y_{k_l+1}-y_{k_l}) - \frac{\gamma}{h}(x_{k_l+1}-x_{k_l}) - \nabla f(x_{k_l}) \in \partial\phi(y_{k_l+1}).$$

Due to the monotoncity of the subdifferential $\partial\phi$, we have

$$\langle -\frac{1}{h}(y_{k_l+1}-y_{k_l}) - \frac{\gamma}{h}(x_{k_l+1}-x_{k_l}) - \nabla f(x_{k_l}) - \partial\phi(0), \frac{y_{k_l+1}}{\|y_{k_l+1}\|}\rangle \geq 0 \quad \forall l \in \mathbb{N}. \tag{2.13}$$

Since the sequence $w_l = \left(\frac{y_{k_l+1}}{\|y_{k_l+1}\|}\right)_l$ is bounded in a finite dimensional space, it is relatively compact, and hence has a convergent subsequence. For notational convenience, we use the same notation and therefore assume that $w_l \to w$. It is clear that $\|w\| = 1$. Let $l \to +\infty$ in (2.13). According to (2.11) and (2.12) since the general term of a convergent series necessarily goes to zero, it follows that

$$\langle \partial\phi(0), w\rangle \leq 0.$$

Since $\overline{\mathbb{B}}(0,r) \subset \partial\phi(0)$, the above inequality implies that

$$\langle ru, w\rangle \leq 0 \quad \forall u \in \overline{\mathbb{B}}(0,1).$$

Choose $u = w$. It follows that $r\|w\|^2 \leq 0$. Therefore $w = 0$, which is in contradiction with $\|w\| = 1$.

2.2 Estimating the transition process

Let us give some information about the number of steps after which the iterates $(x_k)_k$ follow the steepest descent algorithm. According to the proof of Theorem 2.1, this is satisfied as soon as

$$\|z_k\| \leq \frac{hr}{1+\gamma h},$$

where $z_k = \frac{1}{1+\gamma h} y_k + \frac{(\gamma\beta-1)h}{1+\gamma h}\nabla f(x_k)$. Let us take advantage of the summation estimates we have obtained in the proof of Theorem 2.1, that is

$$\sum_{k=1}^{+\infty} \|y_k\| \leq \frac{E_1}{hr}, \quad \sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < \frac{E_1}{h\beta}. \tag{2.14}$$

According to the definition of z_k, elementary algebra gives

$$\|z_k\|^2 \leq \frac{2}{(1+\gamma h)^2}\|y_k\|^2 + \frac{2(\gamma\beta-1)^2h^2}{(1+\gamma h)^2}\|\nabla f(x_k)\|^2.$$

According to (2.14) and the inequality $\sum_{k=1}^{+\infty}\|y_k\|^2 \leq (\sum_{k=1}^{+\infty}\|y_k\|)^2$, we infer

$$\begin{aligned}\sum_{k=1}^{+\infty}\|z_k\|^2 &\leq \frac{2}{(1+\gamma h)^2}\sum_{k=1}^{+\infty}\|y_k\|^2 + \frac{2(\gamma\beta-1)^2h^2}{(1+\gamma h)^2}\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2 \\ &\leq \frac{2}{(1+\gamma h)^2}\left(\frac{E_1}{hr}\right)^2 + \frac{2(\gamma\beta-1)^2h^2}{(1+\gamma h)^2}\frac{E_1}{h\beta}.\end{aligned}$$

Set $M := \frac{2}{(1+\gamma h)^2}\left(\frac{E_1}{hr}\right)^2 + \frac{2(\gamma\beta-1)^2h^2}{(1+\gamma h)^2}\frac{E_1}{h\beta}$. We have

$$\sum_{k=1}^{+\infty}\|z_k\|^2 \geq \sum_{i=k}^{2k}\|z_i\|^2 \geq k \inf_{k\leq i\leq 2k}\|z_i\|^2.$$

Therefore

$$\inf_{k\leq i\leq 2k}\|z_i\| \leq \sqrt{\frac{M}{k}}.$$

Combining the above results, we obtain that

$$k \geq \frac{M(1+h\gamma)^2}{h^2r^2} \Longrightarrow \exists i,\ k \leq i \leq 2k \text{ such that } \frac{1}{h}(x_{i+1}-x_i) + \beta\nabla f(x_i) = 0.$$

Let us now establish the convergence rate of y_k

2.3 Exponential convergence rate of (y_k) to zero

Recall that $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$, $k \geq 1$.

Proposition 2.1 *Set $q = \frac{1}{\sqrt{1+2\gamma h}} \in (0,1)$. Then, there exists $k_0 \in \mathbb{N}$ such that*

$$\|y_k\| \leq q^{k-k_0}\|y_{k_0}\| \quad \forall k > k_0.$$

Proof. The convergence rate of $(y_k)_k$ can be established as follows. First, we have

$$y_{k+1} - y_k + \gamma(x_{k+1} - x_k) + h\partial\phi(y_{k+1}) + h\nabla f(x_k) \ni 0.$$

Taking the scalar product of the above inclusion with y_{k+1}, and using the convexity of ϕ, we obtain

$$\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \gamma\langle x_{k+1} - x_k, y_{k+1}\rangle + h\phi(y_{k+1}) + h\langle\nabla f(x_k), y_{k+1}\rangle \leq 0. \tag{2.15}$$

Since $\nabla f(x_k) \to 0$, we have $(\gamma\beta - 1)\nabla f(x_k) \in \partial\phi(0)$ for k sufficiently large. By definition of the subdifferential, we deduce that

$$\phi(y_{k+1}) \geq (\gamma\beta - 1)\langle\nabla f(x_k), y_{k+1}\rangle.$$

Equivalently

$$\phi(y_{k+1}) + \langle\nabla f(x_k), y_{k+1}\rangle \geq \gamma\beta\langle\nabla f(x_k), y_{k+1}\rangle.$$

According to the above inequality and the Cauchy-Schwarz inequality, from (2.15) we deduce that

$$\frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + \gamma\langle x_{k+1} - x_k, y_{k+1}\rangle + \gamma\beta h\langle\nabla f(x_k), y_{k+1}\rangle \leq 0.$$

Equivalently

$$\frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + \gamma h\langle\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k), y_{k+1}\rangle \leq 0.$$

According to the definition of y_{k+1}, this gives

$$(1 + 2\gamma h)\|y_{k+1}\|^2 \leq \|y_k\|^2.$$

Set $q = \frac{1}{\sqrt{1+2\gamma h}} \in (0,1)$, we finally deduce that $\|y_{k+1}\| \leq q\|y_k\|$ for k sufficiently large, say $k \geq k_0$. Therefore,

$$\|y_k\| \leq q^{k-k_0}\|y_{k_0}\| \quad \forall k > k_0.$$

The proof is thereby completed. ■

3 The convex case

3.1 General convex case

Let us state our main result concerning the convergence properties of the algorithm (IPAHDD-C1) when f is a general convex function.

Theorem 3.1 *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex, differentiable function whose gradient is L-Lipschitz continuous, and such that $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. Assume that the friction potential $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$hL \leq \frac{2\gamma}{\gamma\beta + 1}. \tag{3.1}$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1) *satisfies the following properties:*

(i) *$(x_k)_k$ converges weakly, and its limit is a minimizer of f.*

(ii) *$\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0$ after a finite number of steps.*

(iii) *$\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty$ and $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$.*

Proof. According to Theorem 2.1, after a finite number of steps, say $k \geq k_0$

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

i.e. , the sequence $(x_k)_{k\geq k_0}$ follows the classical gradient scheme with the fixed step size $s = h\beta > 0$. It is then a classical result (see for example [22, Corollary 27.9]) that the sequence converges weakly, and its limit is a minimizer of f, whenever the step size s satisfies

$$s = h\beta < \frac{2}{L}.$$

Clearly this is satisfied, because, under the assumption (3.1) on the parameters, we have

$$hL \leq \frac{2\gamma}{\gamma\beta + 1} < \frac{2\gamma}{\gamma\beta} = \frac{2}{\beta}.$$

Let us recall that the Opial's lemma (stated below) is the key ingredient to prove the weak convergence of the iterates. ■

Lemma 3.1 *Let S be a nonempty set of a Hilbert space $\mathcal{H}$. Suppose that $(x_k)_k$ is a sequence in $\mathcal{H}$ which satisfies*

- *$\lim_{k\to\infty} \|x_k - p\|$ exists for all $p \in S$.*
- *For each subsequence $(x_{k_l})_l$ of $(x_k)_k$ that converges weakly to x, we have $x \in S$.*

Then, there exists $x \in S$ such that $(x_k)_k$ converges weakly to x.

Remark 3.1 In Theorem 3.1, let us give a direct proof that $(x_k)_k$ converges weakly to a minimizer of f, without using the fact that after a finite number of steps, the iterates follow the steepest descent method. The proof is based on Opial's lemmma. According to the convexity of f, and hence the monotonicity of ∇f, we have for all $k \geq 1$ and for all $z \in \mathcal{H}$

$$\begin{aligned}\beta\langle\nabla f(x_{k-1}), x_{k-1} - z\rangle &= \beta\langle\nabla f(x_{k-1}) - \nabla f(z), x_{k-1} - z\rangle + \beta\langle\nabla f(z), x_{k-1} - z\rangle \\ &\geq \beta\langle\nabla f(z), x_{k-1} - z\rangle.\end{aligned}$$

Therefore,

$$\begin{aligned}\langle y_k, x_{k-1}-z\rangle &= \langle \frac{1}{h}(x_k-x_{k-1}), x_{k-1}-z\rangle + \beta\langle \nabla f(x_{k-1}), x_{k-1}-z\rangle \\ &\geq \frac{1}{2h}(\|x_k-z\|^2 - \|x_{k-1}-z\|^2 - \|x_k-x_{k-1}\|^2) + \beta\langle \nabla f(z), x_{k-1}-z\rangle,\end{aligned}$$

where $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$.
This, together with the Cauchy Schwarz inequality, implies

$$\frac{1}{2h}(\|x_k-z\|^2 - \|x_{k-1}-z\|^2) \leq (\|y_k\| + \beta\|\nabla f(z)\|)\|x_{k-1}-z\| + \frac{1}{2h}\|x_k-x_{k-1}\|^2.$$

To check the first item of the Opial's lemma, let us now assume that $z \in \operatorname{argmin}_{\mathcal{H}} f$ which is fixed. As a result, it follows

$$\frac{1}{2h}(\|x_k-z\|^2 - \|x_{k-1}-z\|^2) \leq \|y_k\|\|x_{k-1}-z\| + \frac{1}{2h}\|x_k-x_{k-1}\|^2. \tag{3.2}$$

By summing the above inequalities from $k=1$ to $N \geq 1$, we obtain

$$\frac{1}{2h}(\|x_N-z\|^2 - \|x_0-z\|^2) \leq \sum_{k=1}^{N}\|y_k\|\|x_{k-1}-z\| + \frac{1}{2h}\sum_{k=1}^{N}\|x_k-x_{k-1}\|^2. \tag{3.3}$$

Recall that we have already obtained $\sum_{k=1}^{\infty}\|x_k-x_{k-1}\|^2 < +\infty$ and $\sum_{k=1}^{\infty}\|y_k\| < +\infty$.
Set $P = \sum_{k=1}^{\infty}\|y_k\| \geq 0$, $Q = \sum_{k=1}^{\infty}\|x_k-x_{k-1}\|^2 \geq 0$ and $m_n = \max_{0\leq i\leq n}\|x_i - z\|$. For all $n \geq 1$ and $1 \leq i \leq n$, we deduce from (3.3) that

$$\frac{1}{2h}(\|x_i-z\|^2 - \|x_0-z\|^2) \leq P.m_{i-1} + \frac{1}{2h}Q \leq P.m_n + \frac{1}{2h}Q.$$

It follows that for all $n \geq 1$, we have

$$m_n^2 - \|x_0-z\|^2 \leq 2hPm_n + Q,$$

or

$$m_n^2 - 2hPm_n - \|x_0-z\|^2 - Q \leq 0.$$

The above inequality implies that

$$m_n \leq hP + \sqrt{h^2P^2 + \|x_0-z\|^2 + Q} \quad \forall n \geq 1,$$

which means that $(m_n)_n$ is bounded, and hence the sequence $(\|x_k-z\|)_k$ is bounded. Combining this boundedness property with (3.3), we can easily show that $(\|x_k-z\|)_k$ is a Cauchy sequence in $\mathbb{R}$, and hence converges. We have shown that $(x_k)_k$ fulfills the first item of the Opial's lemma.
Now, we turn to proving that $(x_k)_k$ also satisfies the second item of the Opial's lemma. To this end, take any subsequence $(x_{k_l})_l$ of $(x_k)_k$ and assume that $(x_{k_l})_l$ converges weakly to some $x \in \mathcal{H}$. Since f is convex, we have for all $z \in \mathcal{H}$

$$f(z) \geq f(x_{k_l}) + \langle \nabla f(x_{k_l}), z - x_{k_l}\rangle.$$

Let us pass to the $\liminf$ as $l \to +\infty$ in the above inequality. Since $(\nabla f(x_{k_l}))_l$ converges strongly to 0 and $(x_{k_l})_l$ is bounded, we obtain

$$f(z) \geq \liminf_{l\to\infty} f(x_{k_l}).$$

Moreover, f is weakly lower semicontinuous, so the above inequality gives

$$f(z) \geq f(x).$$

Since z can be taken arbitrarily in $\mathcal{H}$, we deduce that $x \in \operatorname{argmin}_{\mathcal{H}} f$.
With all things considered, we apply the Opial's lemma to deduce that there exists $x_\infty \in \operatorname{argmin}_{\mathcal{H}} f$ such that $(x_k)_k$ converges weakly to x_∞ in $\mathcal{H}$.

3.2 Strongly convex case

Theorem 3.2 *In addition to the assumptions of Theorem 3.1, let us assume that f is strongly convex with parameter $\mu > 0$ and that*

$$\textit{either } h\beta = \frac{1}{L} \textit{ or } h\beta \leq \frac{2}{\mu + L}.$$

Let us denote by x_∞ the unique minimizer of f. Then, we have linear strong convergence of $(x_k)_k$ to x_∞.

Proof. We have shown that, after a finite number of steps, the sequence (x_k) follows the steepest descent method. Specifically,

$$x_{k+1} = x_k - \beta h \nabla f(x_k) \quad \text{for } k \text{ large enough.}$$

Therefore, the conclusion follows from the classical result concerning the convergence rate of the steepest descent method for strongly convex objective functions, see for example [20]. ■

4 The nonconvex case

In this section, $\mathcal{H} = \mathbb{R}^N$ is the finite dimensional Euclidean space. This will allow us to use the Kurdyka–Lojasiewicz property, which we briefly designate by (KL). No convexity assumption is made on the function f to be minimized, which will be assumed to satisfy (KL).

4.1 Some basic facts concerning (KL)

A function $f : \mathbb{R}^N \to \mathbb{R}$ satisfies the (KL) property if its values can be reparametrized in the neighborhood of each of its critical points, so that the resulting function becomes sharp. This means that there exists a continuous, concave, increasing function θ such that for all u in a slice of f

$$\|\nabla(\theta \circ f)(u)\| \geq 1.$$

The function θ captures the geometry of f around its critical points, and is called a desingularizing function; see [8], [7], [25] for further results. Tame functions satisfy the property (KL). Tameness refers to a ubiquitous geometric property of functions and sets encountered in most finite dimensional optimization problems. Sets or functions are called tame when they can be described by a finite number of basic formulas/ inequalities/Boolean operations involving standard functions such as polynomial, exponential, or max functions. Classical examples of tame objects are piecewise linear objects (with finitely many pieces), or semi-algebraic objects. The general notion covering these situations is the concept of o-minimal structure; see van den Dries [33]. Tameness models nonsmoothness via the so-called stratification property of tame sets/functions. It was this property which motivated the vocable of tame topology, la topologie modre according to Grothendieck. All these aspects have been well documented in a series of recent papers devoted to nonconvex nonsmooth optimization, see Ioffe [36], Castera–Bolte–Févotte–Pauwels [30] for an application to deep learning, and [7] for illustrations, examples within a general optimization setting.

4.2 Convergence under (KL) property

Theorem 4.1 *Take $\mathcal{H} = \mathbb{R}^N$. Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and which satisfies the (KL) property. Assume that $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$hL \leq \frac{2\gamma}{\gamma\beta + 1}.$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1) *satisfies the following properties:*

(i) *$(x_k)_k$ converges, and its limit is a critical point of f.*

(ii) *$\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0$ after a finite number of steps.*

(iii) *$\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty$ and $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$.*

Proof. We have shown that, after a finite number of steps, the sequence (x_k) follows the steepest descent method. Specifically,

$$x_{k+1} = x_k - \beta h \nabla f(x_k) \quad \text{for } k \text{ large enough.}$$

Therefore, the conclusion follows from the convergence result of Attouch, Bolte and Svaiter [8, Theorem 3.2] concerning the convergence of the gradient method for functions satisfying the (KL) property. ■

5 Errors, perturbations

Let us examine the effect of introducing perturbations, errors in the algorithm (IPAHDD-C1). According to the dynamic approach, lets start from the perturbed version of (1.1)

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

where the right-hand side $e(\cdot)$ takes into account perturbations, errors. A temporal discretization similar to that in Section 2 gives

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \partial\phi(\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)) \\ + \frac{\beta}{h}(\nabla f(x_k) - \nabla f(x_{k-1})) + \nabla f(x_k) \ni e_k. \qquad (5.1)$$

Solving the above inclusion with respect to x_{k+1} gives the following algorithm:

(IPAHDD-C1-pert)

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

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

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

We have the following convergence results for this perturbed version of the algorithm (IPAHDD-C1).

Theorem 5.1 *Lets make the assumptions of Theorem 2.1, and suppose that the sequence $(e_k)_k$ of perturbations, errors satisfies:*

$$\lim_k \|e_k\| = 0 \quad as \quad k \to +\infty.$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C1-pert) *satisfies the following properties:*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta \nabla f(x_k) = 0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty$.

Proof. The proof is parallel to that of Theorem 2.1. Multiplying both sides of (5.1) with h, and rewriting it using y_k, we obtain for $k \geq 1$

$$y_{k+1} - y_k + \gamma(x_{k+1} - x_k) + h\partial\phi(y_{k+1}) + h\nabla f(x_k) \ni he_k. \tag{5.2}$$

Taking the scalar product of (5.2) with y_{k+1}, we obtain

$$\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \gamma\langle x_{k+1} - x_k, \frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)\rangle + h\langle \partial\phi(y_{k+1}), y_{k+1}\rangle$$
$$+ h\langle \nabla f(x_k), \frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)\rangle = h\langle e_k, y_{k+1}\rangle.$$

According to the assumption (2.3) on the parameters, similar calculation as in Theorem 2.1 gives

$$\frac{1}{2}\|y_{k+1}\|^2 - \frac{1}{2}\|y_k\|^2 + (\gamma\beta + 1)(f(x_{k+1}) - f(x_k)) + h\phi(y_{k+1}) + \beta h\|\nabla f(x_k)\|^2 \leq h\|e_k\|\|y_{k+1}\|.$$

Equivalently

$$E_{k+1} - E_k + h\phi(y_{k+1}) + \beta h\|\nabla f(x_k)\|^2 \leq h\|e_k\|\|y_{k+1}\|, \tag{5.3}$$

where

$$E_k := \frac{1}{2}\|y_k\|^2 + (\gamma\beta + 1)\left(f(x_k) - \inf_{x\in H} f(x)\right).$$

Since ϕ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$, we deduce that

$$E_{k+1} - E_k + h(r - \|e_k\|)\|y_{k+1}\| + \beta h\|\nabla f(x_k)\|^2 \leq 0.$$

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

$$E_{k+1} - E_k + \frac{hr}{2}\|y_{k+1}\| + \beta h\|\nabla f(x_k)\|^2 \leq 0.$$

By summing the above inequalities we deduce that

$$\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty, \text{ and } \sum_{k=1}^{+\infty} \|y_k\| < +\infty. \tag{5.4}$$

Let us now prove that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent method. The proof relies on Lemma 1.1. Recall that, according to (5.2), we have the following equivalent formulation of the algorithm (IPAHDD-C1-pert)

$$y_{k+1} = \mathrm{prox}_{\frac{h}{1+\gamma h}\phi}(z_k),$$

where

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

According to (5.4), since the general term of a convergent series necessarily goes to zero, we have that

$$\lim_k \nabla f(x_k) = \lim_k y_k = 0.$$

Consequently, according to the definition of z_k, and since e_k tends to zero, we have $\lim_k z_k = 0$. Therefore, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\|z_k\| \leq \frac{hr}{1+\gamma h}.$$

According to Lemma 1.1, this implies that $y_{k+1} = \operatorname{prox}_{\frac{h}{1+\gamma h}\phi}(z_k) = 0$ for all $k \geq k_0$. Equivalently, $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0$ which means that after a finite number of steps, the sequence (x_k) follows the steepest descent algorithm. This completes the proof. ■

Remark 5.1 As an immediate consequence of Theorem 5.1, we obtain the convergence of the sequence (x_k) in the perturbed convex case, and in the perturbed nonconvex case under (KL).

Remark 5.2 For the Nesterov accelerated gradient method, which is based on an inertial dynamic with asymptotic vanishing viscous friction, introducing errors e_k does not affect the fast convergence property as long as $\sum_k k\|e_k\| < +\infty$. By contrast, in our situation, to preserve the convergence properties, we just need to assume that $\lim_k \|e_k\| = 0$. This is a remarkable property which is specific to the dry friction damping, and which makes this type of algorithm attractive to deal with noisy/stochastic data.

6 Combining with Nesterov acceleration method

We construct algorithms, still obtained by temporal discretizations of the differential inclusion

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

and which have an analogous structure to the accelerated gradient method of Nesterov [40, 41]. Specifically, we consider the following discretization of the dynamics

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma}{h}(x_{k+1} - x_k) + \partial\phi(\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k)) \\ + \frac{\beta}{h}(\nabla f(x_k) - \nabla f(x_{k-1})) + \nabla f(z_k) \ni 0. \qquad (6.1)$$

There is some flexibility in the choice of the point z_k where the gradient of f is computed. By taking $z_k = x_k$, we obtain the algorithm (IPAHDD-C1) studied in section 2. In this section, we consider two different choices for z_k, which are in accordance with the Nesterov accelerated gradient method:

6.1 Case 1

Take $z_k = x_k + \frac{1}{1+\gamma h}(x_k - x_{k-1})$. With this choice of z_k in (6.1), elementary calculation gives the following algorithm:

(IPAHDD-C2)

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

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

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

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

Theorem 6.1 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the friction potential function $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$\begin{cases} \gamma > \max\{2hL, L/2\}, \\ \beta < \min\left\{\frac{\gamma+\gamma^2 h-2Lh}{Lh}, \frac{2+(2\gamma-L)h}{\gamma^2 h+\gamma}\right\}. \end{cases}$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C2) *satisfies the following properties:*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{+\infty} \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 < +\infty$.

Proof. Let us rewrite (6.1) with the help of $y_k = \frac{1}{h}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1})$. Equivalently, we have

$$y_{k+1} - y_k + \gamma(x_{k+1} - x_k) + h\partial\phi(y_{k+1}) + h\nabla f(z_k) \ni 0.$$

By taking the scalar product of the above inclusion with y_{k+1} we obtain

$$\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \gamma\langle x_{k+1} - x_k, y_{k+1}\rangle + h\langle\partial\phi(y_{k+1}), y_{k+1}\rangle + h\langle\nabla f(z_k), y_{k+1}\rangle = 0. \tag{6.2}$$

We can easily check that

$$\gamma\langle x_{k+1} - x_k, y_{k+1}\rangle = \frac{\gamma h}{2}\|y_{k+1}\|^2 + \frac{\gamma}{2h}\|x_{k+1} - x_k\|^2 - \frac{\gamma h\beta^2}{2}\|\nabla f(x_k)\|^2. \tag{6.3}$$

According to the L-Lipschitz continuity of ∇f, we also have

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

Moreover, according to the gradient descent lemma

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

By combining the two estimates above, we obtain

$$\begin{aligned} h\langle \nabla f(z_k), y_{k+1}\rangle \geq \frac{-h^2L}{1+\gamma h}\|y_k\|\|y_{k+1}\| - \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_{k-1})\|^2 + \|y_{k+1}\|^2) \\ + \beta h\|\nabla f(x_k)\|^2 + f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1}-x_k\|^2. \end{aligned} \tag{6.4}$$

By combining (6.2), (6.3) and (6.4), and using the dry friction property $\phi(u) \geq r\|u\|$, we obtain

$$\begin{aligned} &\|y_{k+1}\|^2 - \langle y_k, y_{k+1}\rangle + \frac{\gamma h}{2}\|y_{k+1}\|^2 + \frac{\gamma}{2h}\|x_{k+1}-x_k\|^2 - \frac{\gamma h\beta^2}{2}\|\nabla f(x_k)\|^2 + hr\|y_{k+1}\| \\ &- \frac{h^2L}{1+\gamma h}\|y_k\|\|y_{k+1}\| - \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_{k-1})\|^2 + \|y_{k+1}\|^2) \\ &+ \beta h\|\nabla f(x_k)\|^2 + f(x_{k+1}) - f(x_k) - \frac{L}{2}\|x_{k+1}-x_k\|^2 \leq 0. \end{aligned}$$

Therefore,

$$\begin{aligned} &(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_{k+1}\|^2 - (1 + \frac{h^2L}{1+\gamma h})\|y_k\|\|y_{k+1}\| + (\frac{\gamma}{2h} - \frac{L}{2})\|x_{k+1}-x_k\|^2 \\ &+ (\beta h - \frac{\gamma h\beta^2}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|\nabla f(x_k)\|^2 + \frac{h^2L\beta}{2(1+\gamma h)}(\|\nabla f(x_k)\|^2 - \|\nabla f(x_{k-1})\|^2) \\ &+ f(x_{k+1}) - f(x_k) + hr\|y_{k+1}\| \leq 0. \end{aligned}$$

For each $k \geq 1$ set

$$E_k := \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_k\|^2 + \frac{h^2L\beta}{2(1+\gamma h)}\|\nabla f(x_{k-1})\|^2 + f(x_k) - \inf_{\mathcal{H}} f. \tag{6.5}$$

We deduce that

$$\begin{aligned} &E_{k+1} - E_k + \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_{k+1}\|^2 - (1 + \frac{h^2L}{1+\gamma h})\|y_k\|\|y_{k+1}\| \\ &+ \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_k\|^2 + (\frac{\gamma}{2h} - \frac{L}{2})\|x_{k+1}-x_k\|^2 \\ &+ (\beta h - \frac{\gamma h\beta^2}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|\nabla f(x_k)\|^2 + hr\|y_{k+1}\| \leq 0. \end{aligned}$$

According to the assumptions on γ, h and β, we have

$$\begin{cases} \frac{\gamma}{2h} - \frac{L}{2} \geq 0, \\ \beta h - \frac{\gamma h\beta^2}{2} - \frac{h^2L\beta}{2(1+\gamma h)} > 0. \end{cases}$$

Let us show that

$$\frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_{k+1}\|^2 - (1 + \frac{h^2L}{1+\gamma h})\|y_k\|\|y_{k+1}\| + \frac{1}{2}(1 + \frac{\gamma h}{2} - \frac{h^2L\beta}{2(1+\gamma h)})\|y_k\|^2 \geq 0.$$

Indeed, a sufficient condition for this is

$$\begin{cases} 1+\frac{\gamma h}{2}-\frac{h^2L\beta}{2(1+\gamma h)}>0, \\ (1+\frac{h^2L}{1+\gamma h})^2-(1+\frac{\gamma h}{2}-\frac{h^2L\beta}{2(1+\gamma h)})^2\leq 0. \end{cases}$$

Equivalently (since $\gamma>0, h>0, \beta>0$)

$$1+\frac{h^2L}{1+\gamma h}\leq 1+\frac{\gamma h}{2}-\frac{h^2L\beta}{2(1+\gamma h)},$$

or

$$\beta\leq\frac{\gamma+\gamma^2h-2Lh}{Lh}, \tag{6.6}$$

which is fullfiled, according to our assumptions on γ, h and β.

We have shown that

$$E_{k+1}-E_k+(\beta h-\frac{\gamma h\beta^2}{2}-\frac{h^2L\beta}{2(1+\gamma h)})\|\nabla f(x_k)\|^2+hr\|y_{k+1}\|\leq 0, \tag{6.7}$$

where E_k has been defined in (6.5). By summing the above inequalities, we obtain

$$\sum_{k=1}^{+\infty}\|\nabla f(x_k)\|^2<+\infty, \quad \sum_{k=1}^{+\infty}\|y_k\|<+\infty. \tag{6.8}$$

Let us now prove that after a finite number of steps, the sequence $(x_k)_k$ follows the steepest descent method. The proof relies on Lemma 1.1. Recall the following equivalent formulation of (IPAHDD-C2)

$$y_{k+1}=\operatorname{prox}_{\frac{h}{1+\gamma h}\phi}(w_k),$$

where

$$w_k=\frac{1}{h}(z_k-x_k)+\frac{\beta}{1+\gamma h}\nabla f(x_{k-1})+\frac{h\beta\gamma}{1+\gamma h}\nabla f(x_k)-\frac{h}{1+\gamma h}\nabla f(z_k).$$

According to (6.8), and since the general term of a convergent series necessarily goes to zero, we have that

$$\lim_k \nabla f(x_k)=\lim_k y_k=0.$$

By definition of y_k this implies

$$\lim_k x_k-x_{k-1}=0.$$

According to the Lipschitz continuity of ∇f, we easily deduce that $\lim_k w_k=0$. Therefore, there exists $k_0\in\mathbb{N}$ such that for all $k\geq k_0$,

$$\|w_k\|\leq\frac{hr}{1+\gamma h}.$$

According to Lemma 1.1, this implies that $y_{k+1}=\operatorname{prox}_{\frac{h}{1+\gamma h}\phi}(w_k)=0$ for all $k\geq k_0$. Equivalently, $\frac{1}{h}(x_{k+1}-x_k)+\beta\nabla f(x_k)=0$ which means that after a finite number of steps, the sequence (x_k) follows the steepest descent algorithm. This completes the proof. ■

6.2 Case 2

Take $z_k = x_k + \frac{1}{h(1+\gamma h)}(x_k - x_{k-1})$ in (6.1). With this choice of z_k, elementary calculation gives the following algorithm:

(IPAHDD-C3)

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

$$z_k = x_k + \tfrac{1}{h(1+\gamma h)}(x_k - x_{k-1}).$$

$$w_k = z_k - x_k + \tfrac{\beta}{1+\gamma h}\nabla f(x_{k-1}) + \tfrac{h\beta\gamma}{1+\gamma h}\nabla f(x_k) - \tfrac{h}{1+\gamma h}\nabla f(z_k)$$

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

A similar proof to the one of Theorem 6.1 gives

Theorem 6.2 *Let $f : \mathcal{H} \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and such that $\inf_{\mathcal{H}} f > -\infty$. Assume that the friction potential function $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters h, γ, β satisfy the relation*

$$\begin{cases} \gamma > \max\left\{\frac{L}{2h}, 2L, Lh\right\}, \\ \beta < \min\left\{\frac{2+2\gamma h - L}{\gamma(1+\gamma h)}, \frac{\gamma + h\gamma^2 - 2L}{L}\right\}. \end{cases}$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C3) *satisfies the following properties:*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta\nabla f(x_k) = 0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{\infty} \|\nabla f(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|^2 < +\infty$.

Remark 6.1 As an immediate consequence of Theorem 6.1 and 6.2, and of the classical properties of the steepest descent method, we obtain the convergence of the sequence (x_k) in the convex case, and in the nonconvex case under (KL). Similar results are still valid for the perturbed version of these algorithms, just assuming that the perturbation terms go to zero asymptotically.

Remark 6.2 In Theorems 6.1 and 6.2, a crucial assumption is $\gamma > \max\{2hL, L/2\}$, resp. $\gamma > \max\{L/2h, 2L, Lh\}$. Thus the viscous damping coefficient γ has to remain large enough. So, the above approach excludes the case where the viscous damping asymptotically goes to zero, which is the case of the Nesterov accelerated gradient method. It is an open question to develop our analysis to cover this situation.

7 Nonsmooth problems

We consider the extension of our study to two nonsmooth situations: the nonsmooth convex case, and the nonsmooth d.c. optimization.

7.1 Nonsmooth convex case

Suppose that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a closed, convex and proper function such that $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. We will reduce to the smooth case by means of the Moreau-Yosida approximation of f. Recall that the Moreau envelope of f of index $\lambda > 0$ is the function $f_\lambda : \mathcal{H} \to \mathbb{R}$ defined by, for all $x \in \mathcal{H}$,

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

As a classical result, f_λ is convex, differentiable and its gradient is $\frac{1}{\lambda}$ Lipschitz continuous. Moreover, we have $\operatorname{argmin}_{\mathcal{H}} f = \operatorname{argmin}_{\mathcal{H}} f_\lambda$ and $\min_{\mathcal{H}} f = \min_{\mathcal{H}} f_\lambda$. One can consult [10, 22, 29] for an in-depth study of the properties of the Moreau envelope in a Hilbert framework. Exploiting this property of the Moreau envelope, we can equivalently consider the problem in which f is substituted by its Moreau envelope, and hence we recover the smooth case. Since $\nabla f_\lambda(x) = \frac{1}{\lambda}(x - \operatorname{prox}_{\lambda f}(x))$, we obtain the following algorithm:

(IPAHDD-C-nonsmooth)

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

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

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 the algorithm (IPAHDD-C-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 (IPAHDD-C-nonsmooth).

Theorem 7.1 *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a closed, convex, proper function such that $\operatorname{argmin}_{\mathcal{H}} f \neq \emptyset$. Assume that the friction potential function $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Suppose that the positive parameters $h, \gamma, \beta, \lambda$ satisfy the relation*

$$\frac{\gamma}{h} - \frac{1}{2\lambda}(\gamma\beta + 1) \geq 0.$$

Then any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C-nonsmooth) *converges weakly and its limit is a minimizer of f. Moreover,*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \frac{\beta}{\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_k)) = 0$ *after a finite number of steps;*

(ii) $\sum_{k=1}^{+\infty} \|x_k - \operatorname{prox}_{\lambda f}(x_k)\|^2 < +\infty$.

Proof. By replacing the Lipschitz constant L in Theorem 2.1 by $\frac{1}{\lambda}$, and using the equality $\nabla f_\lambda(x_k) = \frac{1}{\lambda}(x_k - \operatorname{prox}_{\lambda f}(x_k))$, the result follows immediately. ■

7.2 Nonsmooth nonconvex d.c. problems

Suppose that $f = g - h$ where $g, h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are closed, convex and proper functions. Following Hiriart-Urruty [35], consider the problem in which f is substituted by the difference of the Moreau envelopes of g and h, so recovering the smooth case. Given $\lambda > 0$, according to the properties of the Moreau envelope, the regularized function $\psi_\lambda : \mathcal{H} \to \mathbb{R}$ defined by

$$\psi_\lambda = g_\lambda - h_\lambda,$$

is differentiable and its gradient is $\frac{2}{\lambda}$ Lipschitz continuous. Moreover, if x is a critical point of ψ_λ, we have

$$\begin{aligned} \nabla\psi_\lambda(x) &= \nabla g_\lambda(x) - \nabla h_\lambda(x) \\ &= -\frac{1}{\lambda}\left(\operatorname{prox}_{\lambda g}(x) - \operatorname{prox}_{\lambda h}(x)\right) = 0. \end{aligned}$$

Therefore, $u := \mathrm{prox}_{\lambda g}(x) = \mathrm{prox}_{\lambda h}(x)$, and the point u, which is so defined, verifies $\partial g(u) - \partial h(u) \ni 0$, which is a critical point of $f = g - h$ in the sense of Toland [49]. The algorithm now writes

(IPAHDD-CDC)

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

$$y_k = \tfrac{1}{h}(x_k - x_{k-1}) - \tfrac{\beta}{\lambda}(\mathrm{prox}_{\lambda g}(x_{k-1}) - \mathrm{prox}_{\lambda h}(x_{k-1}))$$

$$x_{k+1} = x_k + \tfrac{\beta h}{\lambda}(\mathrm{prox}_{\lambda g}(x_k) - \mathrm{prox}_{\lambda h}(x_k)) + h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\phi}\left(\tfrac{1}{1+\gamma h}y_k - \tfrac{(\gamma\beta-1)h}{(1+\gamma h)\lambda}(\mathrm{prox}_{\lambda g}(x_k) - \mathrm{prox}_{\lambda h}(x_k))\right)$$

According to the above results, a direct adaptation of Theorem 2.1 gives the following result:

Theorem 7.2 *Let $f = g - h$ where $g, h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are closed, convex and proper functions. Assume that the friction potential $\phi : \mathcal{H} \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Take $\lambda > 0$, and suppose that the positive parameters h, γ, β satisfy the relation*

$$\frac{h}{\lambda} \leq \frac{\gamma}{\gamma\beta + 1}. \tag{7.1}$$

Then, for any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-CDC)*, we have that $(x_k)_k$ satisfies*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta(\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k)) = 0$ *after a finite number of steps.*

(ii) $\sum_{k=1}^{\infty} \|\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k)\|^2 < +\infty$ *and* $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|^2 < +\infty$.

(iii) *If $\mathcal{H}$ is a finite dimensional space, and $g_\lambda - h_\lambda$ verifies the (KL) property, then the sequence (x_k) converges to some x_∞ such that $u := \mathrm{prox}_{\lambda g}(x_\infty) = \mathrm{prox}_{\lambda h}(x_\infty)$ is a critical point in the sense of Toland of $f = g - h$, i.e. ,*

$$\partial g(u) - \partial h(u) \ni 0.$$

Remark 7.1 As a particular case of practical importance, suppose that g and h are convex functions which are semialgebraic. Then their Moreau envelopes are still semialgebraic [8], and so is the difference of their Moreau envelopes. In this case, we have that $g_\lambda - h_\lambda$ verifies the (KL) property, and so the above convergence result is valid in this nonsmooth nonconvex situation.

8 Splitting algorithms for the Lasso-type problems

Take $\mathcal{H} = \mathbb{R}^n$. We consider Lasso-type splitting algorithms for additively structured minimization problems. The function f to be minimized is written

$$f(x) = \tfrac{1}{2}\|Ax - b\|^2 + g(x),$$

where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed, convex proper function.
A direct application of the nonsmooth algorithm (IPAHDD-C-nonsmooth) to this minimization problem would require calculating (at least approximately) the proximal operator of f. Its not easy in general. To overcome this difficulty, we use a change of metric, a technique already used in [1], [13]. 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 closed, convex function f, the Moreaus 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 $x \in \mathbb{R}^n$,

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

The Moreau envelope f_λ^M is a smooth function whose gradient for the Euclidean structure is given by

$$\nabla f_\lambda^M(x) = \frac{1}{\lambda} M(x - \operatorname{prox}_{\lambda f}^M(x)), \tag{8.1}$$

where $\operatorname{prox}_{\lambda f}^M(x) = \operatorname{argmin}_{\xi\in H} \left\{ f(\xi) + \frac{1}{2\lambda}\|x-\xi\|_M^2 \right\}$. 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

$$\|\nabla f_\lambda^M(x_1) - \nabla f_\lambda^M(x_2)\| \leq \frac{1}{\lambda}\sqrt{\frac{\mu_{\max}(M)}{\mu_{\min}(M)}}\|x_1 - x_2\|.$$

We set $M = I_n - \lambda A^T A$. If $\lambda \in [0, \frac{1}{\|A\|^2})$, then M is positive definite. In this case, we have

$$\operatorname{prox}_{\lambda f}^M(x) = \operatorname{prox}_{\lambda g}(x - \lambda A^T(Ax-b)). \tag{8.2}$$

The formulation (8.2) can be consulted in [31]. Using (8.1) and (8.2), we get

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

Since $\operatorname{argmin}_{\mathcal{H}} f_\lambda^M = \operatorname{argmin}_{\mathcal{H}} f$, we can replace f with f_λ^M to recover the smooth case, and obtain

(IPAHDD-C-lasso)

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

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

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

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

Theorem 8.1 *Let A be an $m \times n$ matrix, $b \in \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed, convex proper function. Take $f = \frac{1}{2}\|A\cdot - b\|^2 + g$ and suppose that $\operatorname{argmin}_{\mathbb{R}^n} f \neq \emptyset$. Assume that $\phi : \mathbb{R}^n \to \mathbb{R}$ satisfies the dry friction property $(\mathrm{DF})_r$ for some $r > 0$. Set $M = I_n - \lambda A^T A$ with $\lambda \in [0, \frac{1}{\|A\|^2}[$, and suppose that the positive parameters $h, \gamma, \beta, \lambda$ satisfy the relation*

$$\tfrac{\gamma}{h} - \tfrac{1}{2\lambda}\sqrt{\tfrac{\mu_{max}(M)}{\mu_{min}(M)}}(\gamma\beta + 1) \geq 0.$$

Then, for any sequence $(x_k)_k$ generated by the algorithm (IPAHDD-C-lasso)*, we have that $(x_k)_k$ converges, and its limit is a minimizer of f. Moreover*

(i) $\frac{1}{h}(x_{k+1} - x_k) + \beta z_k = 0$ *after a finite number of steps;*

(ii) $\sum_{k=1}^{\infty} \|z_k\|^2 < +\infty$, *where* $z_k = \frac{1}{\lambda} M(x - \operatorname{prox}_{\lambda g}(x - \lambda A^T(Ax-b)))$.

Proof. Replacing the Lipschitz constant L in Theorem 2.1 by $\frac{1}{\lambda}\sqrt{\mu_{\max}(M)/\mu_{\min}(M)}$, and recalling that $z_k = \nabla f_\lambda^M(x_k)$, then the result follows immediately. ■

9 Some numerical experiments

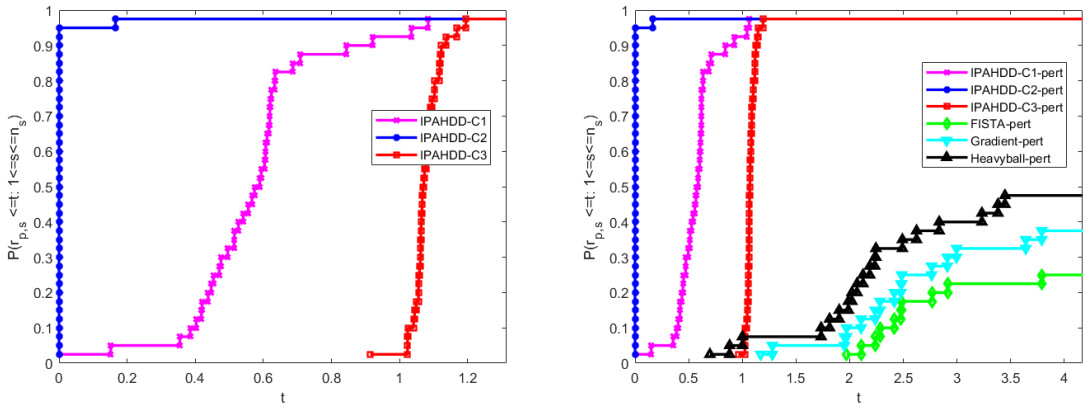
We use the performance profiles developed by Dolan and Moré as a tool for comparing different solvers. For each $t \in \mathbb{R}$, the performance profiles give the proportion $\rho_s(t)$ of test problems on which each solver s under comparison has a performance within the factor t of the best possible ratio. For more details, we refer to [32]. We choose the number of iterations found by each solver as a performance measure.

9.1 Comparing the three algorithms (IPAHDD-C1), (IPAHDD-C2) and (IPAHDD-C3)

We perform numerical tests to compare the algorithms defined in the previous sections, and which deal with general differentiable function f with Lipschitz continuous gradient. We take $\phi : \mathbb{R}^n \to \mathbb{R}$ given by $x \mapsto \phi(x) = r\|x\|$, $r = 0.1$. First consider the simple situation where the function $f : \mathbb{R}^n \to \mathbb{R}$ is quadratic

$$f(x) = \tfrac{1}{2}\|Ax - b\|^2, \quad A \in \mathbb{R}^{m\times n},\ (m \leq n),\ b \in \mathbb{R}^m \text{ are chosen randomly.}$$

The matrices A are generated randomly. We have chosen a set P of 40 different problems with 40 matrices $A \in \mathbb{R}^{m\times n}$. The numerical experiments are carried out on an ordinary computer. All the codes are written and executed in MATLAB R2019a. We use the same initial points and the same stopping criterion, i.e., either the number of iterations exceeds 10^5 or $\|\nabla f(x_k)\| \leq 10^{-6}$. Figure 1(a) reveals that (IPAHDD-C2) is the most efficient method out of the three in the sense that it requires the least number of iterations to reach a solution. Despite their good convergence properties, the algorithms which are based on the dry friction damping are not as fast as the FISTA method. This is easily understandable since our methods are proved to follow the steepest descent method regime after a finite number of steps. However, the situation is reversed if we introduce errors perturbations in the algorithms, as shown in the following experiments.



(a) (b)

Figure 1: Performance profiles of (IPAHDD-C1), (IPAHDD-C2) and (IPAHDD-C3) (left), (IPAHDD-C1-pert), (IPAHDD-C2-pert), (IPAHDD-C3-pert), (FISTA-pert), (Gradient-pert) and (Heavy Ball-pert) (right).

9.2 Introducing errors

Recall that for the heavy ball method, introducing errors (e_k) does not affect the fast convergence property as long as $\sum \|e_k\| < +\infty$. For the FISTA algorithm, the condition is even more stringent, we need to

assume that $\sum k\|e_k\| < +\infty$, see [14, Theorem 5.1] and [45]. A unified presention of these results is given in [12, Theorem 2.1]. By contrast, in our situation, to preserve the convergence properties, we just need to assume that $\lim_k \|e_k\| = 0$. For the development of perturbations aspects of first order optimization methods, interested readers can consult [12,14,17–19,21,24,45,47,50], and [15] in the case of the Hessian driven damping. We will now compare the perturbed versions of our algorithms, namely (IPAHDD-C1-pert), (IPAHDD-C2-pert) and (IPAHDD-C3-pert) (the two latter are respectively the perturbed version of (IPAHDD-C2) and (IPAHDD-C3) and defined in the same way as (IPAHDD-C1-pert)) with the perturbed gradient method, the perturbed Heavy Ball method and the perturbed FISTA method which are given below

(Gradient-pert)
$$\begin{array}{|l} \text{Initialize : } x_0 \in \mathbb{R}^n. \\ x_k = x_{k-1} - \gamma(\nabla f(x_{k-1}) + e_k). \end{array}$$

(Heavy Ball-pert)
$$\begin{array}{|l} \text{Initialize : } y_0 = x_0 \in \mathbb{R}^n. \\ x_{k+1} = x_k + \alpha(x_k - x_{k-1}) - \gamma(\nabla f(x_k) + e_k). \end{array}$$

(FISTA-pert)
$$\begin{array}{|l} \text{Initialize : } y_0 = x_0 \in \mathbb{R}^n, (t_k)_{k\geq 1} : t_k = \frac{k+1}{2}. \\ x_k = y_{k-1} - \gamma(\nabla f(y_{k-1}) + e_k). \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}). \end{array}$$

The sequence $(t_k)_k$ in the above algorithm satisfies $t_1 = 1$ and $t_k^2 \geq t_{k+1}^2 - t_{k+1}$. Under this property, Beck and Teboulle [23] showed the $O(1/k^2)$ convergence rate for the above algorithm in the error-free case, i.e. when $e_k = 0$, $\forall k \geq 1$. Indeed, as explained above, under the summability property $\sum k\|e_k\| < +\infty$, the convergence rate is as in the error-free case (see [14] or [45]). For numerical purposes, we choose the sequence (e_k) such that $\|e_k\| = 1/k$; in fact, for each k we choose a random vector $\xi \in \mathbb{R}^n$ with the uniform distribution on $]0,1[^n$ and then set $e_k = (1/(k\|\xi\|))\xi$. In this way, the conditions $\sum k\|e_k\| < +\infty$ and $\sum \|e_k\| < +\infty$ are not satisfied, which allows us to check the advantage of our methods in presence of perturbations compared to (FISTA-pert), (Gradient-pert) and (Heavy Ball-pert). We use performance profiles on the quadratic problem, as we did before to carry out this comparison. As anticipated, we can see from Figure 1(b) that FISTA, the gradient method and the Heavy Ball method suffer substantially from the errors/perturbations when the conditions $\sum k\|e_k\| < +\infty$ and $\sum \|e_k\| < +\infty$ are not satisfied, while the proposed algorithms prove their robustness and preserve their behavior as in the non-perturbed case. This naturally leads to considering stochastic versions of our algorithms.

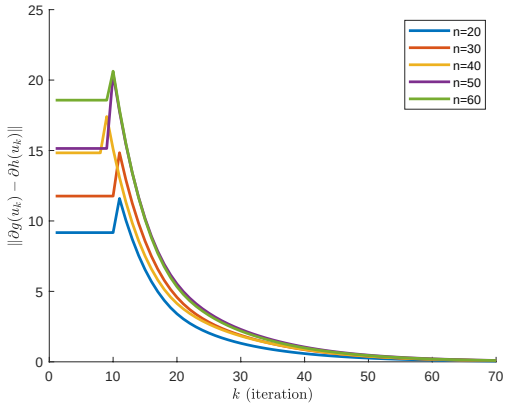
9.3 Nonsmooth nonconvex d.c. problems

Let us illustrate the algorithm (IPAHDD-CDC) with nonsmooth nonconvex problems of DC type. Given $n \geq 2$, consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

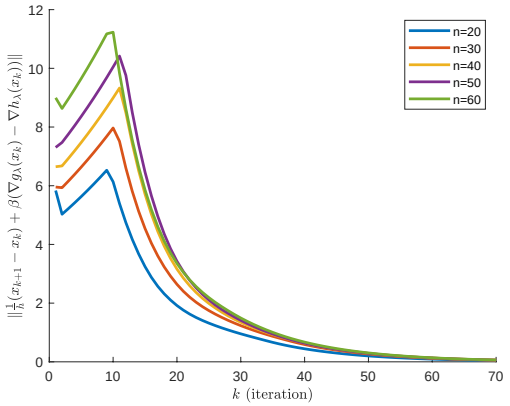
$$f(x) = \|Ax - b\|_2^2 - \|A^T b\|_2 \|x\|_2, \tag{9.1}$$

where A is an orthogonal matrix of order n and $b \in \mathbb{R}^n$. We choose 5 random orthogonal matrices A of size ranging from 20 to 60 while b has all its coordinates equal to one. To apply the algorithm (IPAHDD-CDC), we rely on the "trivial" D.C decomposition $f = g - h$ where $g : x \mapsto \|Ax - b\|_2^2$ and $h : x \mapsto \|A^T b\|_2 \|x\|_2$. Clearly, g and h are semialgebraic. The orthogonality of A is assumed only to

facilitate the computations of prox_g. Therefore, according to Remark 7.1, we have that $g_\lambda - h_\lambda$ satisfies the (KL) property for $\lambda > 0$. As a result, under the assumptions of Theorem 7.2, the sequence (x_k) generated by the algorithm (IPAHDD-CDC) converges to some x_∞, and $\mathrm{prox}_{\lambda h}(x_\infty)$ is a critical point of f in the sense of Toland. It is easy to show that u is critical point of f in the sense of Toland if and only if $u \neq 0$ and $2A^T(Au-b) - \frac{\|A^Tb\|_2 u}{\|u\|_2} = 0$. The stopping condition we use for (IPAHDD-CDC) is either the number of iterations exceeding 10^5 or $u_k \neq 0$ and $\left\|2A^T(Au_k - b) - \frac{\|A^Tb\|_2 u_k}{\|u_k\|_2}\right\|_2 \leq 10^{-6}$. Figure (2) depicts



(a) Behavior of $\|\partial g(u_k) - \partial h(u_k)\|$



(b) Behavior of $\left\|\frac{1}{h}(x_{k+1}-x_k)+\beta(\nabla g_\lambda(x_k)-\nabla h_\lambda(x_k))\right\|$

Figure 2: Algorithm (IPAHDD-CDC) with $f : x \mapsto \|Ax-b\|_2^2 - \|A^Tb\|_2\|x\|_2$ and different initial data.

the behavior of the quantities $\|\partial g(u_k) - \partial h(u_k)\|$ and $\left\|\frac{1}{h}(x_{k+1}-x_k) + \beta(\nabla g_\lambda(x_k) - \nabla h_\lambda(x_k))\right\|$ over iterations, where $u_k = \mathrm{prox}_{\lambda h}(x_k)$, in five problems of different sizes. (IPAHDD-CDC) deals with the five problems successfully. In Figure 2(b), we observe that after a certain number of iterations, the norm of the sum of the discrete velocity vector and gradient terms is decreasing. This is in accordance with Theorem 7.2, which establishes that after a finite number of iterations, the algorithm follows the steepest descent regime. We now consider the algorithm (IPAHDD-CDC-pert) which is a perturbed version of (IPAHDD-CDC).

(IPAHDD-CDC-pert)

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

$$y_k = \tfrac{1}{h}(x_k - x_{k-1}) - \tfrac{\beta}{\lambda}(\mathrm{prox}_{\lambda g}(x_{k-1}) - \mathrm{prox}_{\lambda h}(x_{k-1}))$$

$$x_{k+1} = x_k + \tfrac{\beta h}{\lambda}(\mathrm{prox}_{\lambda g}(x_k) - \mathrm{prox}_{\lambda h}(x_k)) + h\,\mathrm{prox}_{\frac{h}{1+\gamma h}\phi}\left(\tfrac{1}{1+\gamma h}y_k - \tfrac{(\gamma\beta-1)h}{(1+\gamma h)\lambda}(\mathrm{prox}_{\lambda g}(x_k) - \mathrm{prox}_{\lambda h}(x_k) + \tfrac{h}{1+\gamma h}e_k\right)$$

It is easy to check that under the assumptions of Theorem 7.2 together with $\lim_k \|e_k\| = 0$, the conclusions of Theorem 7.2 also hold true for the algorithm (IPAHDD-CDC-pert). It is well-known that the classical DC algorithm (DCA), introduced by Pham Dinh Tao [43] is one of the algorithms that solve effectively nonsmooth and nonconvex optimization problems of the form

$$\inf_{x\in\mathbb{R}^n}\{f(x) := g(x) - h(x)\},$$

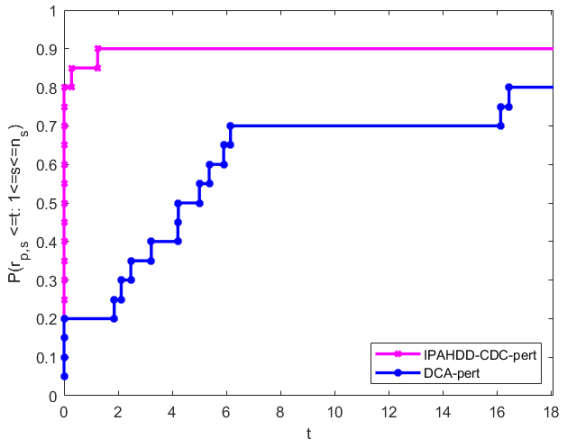


Figure 3: Performance profiles of (IPAHDD-CDC-pert) and (DCA-pert) on the problem (9.1)

where g and h are lower semicontinuous proper real extended valued convex functions. Briefly, the algorithm consists in constructing two sequences (x_k) and (y_k) such that the sequences of values of the primal and dual objective functions $\{g(x_k) - h(x_k)\}, \{g^*(x_k) - h^*(x_k)\}$ are decreasing, and their corresponding limits x_∞ and y_∞ satisfy local optimality conditions [38]. Precisely, the standard (DCA) reads as follows. Choose an initial point $x_0 \in \text{dom}(g) = \{x \in \mathbb{R}^n \ : \ g(x) < +\infty\}$, and for $k = 0, 1, \ldots,$ set

$$y_k \in \partial h(x_k); \quad x_{k+1} \in \partial g^*(y_k) = \text{argmin}_{x \in \mathbb{R}^n} \left\{ g(x) - \langle y^k, x \rangle \right\}.$$

For the purpose of comparison with (IPAHDD-CDC-pert), we propose the perturbed version of DCA

(DCA-pert)

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

$y_k \in \partial h(x_k)$

$x_{k+1} \in \partial g^*(y_k) + e_k = \text{argmin}_{x \in \mathbb{R}^n} \left\{ g(x) - \langle y^k, x \rangle \right\} + e_k$

Using performance profiles with the number of iterations as a performance measure, we make a comparison between (IPAHDD-CDC-pert) and (DCA-pert) on the d.c problem (9.1). The perturbation sequence here is chosen in the same way as before, i.e., for each k we choose a random vector $\xi \in \mathbb{R}^n$ with the uniform distribution on $]0, 1[^n$ and then set $e_k = \frac{\xi}{k\|\xi\|}$. The performance profiles in Fig. 3 show that in the presence of perturbations, (IPAHDD-CDC-pert) outperforms (DCA-pert). Specifically, (IPAHDD-CDC-pert) wins over (DCA-pert) on 80% of the problems used for this experiment; moreover, the number of problems that can be solved by (IPAHDD-CDC-pert) is higher (compared to (DCA-pert)).

10 Concluding remarks

In this paper, we presented a new way of handling dry friction in first order inertial algorithms. While in previous works, dry friction comes as a nonlinear action on the velocity, we now consider its action on a weighted sum of the velocity vector and the gradient of the function f to be minimized. As a first favorable property, the sequences thus generated converge towards critical points of f (global minima when f is convex), whereas previously we only end up with approximate critical points of f. In addition, after a finite number of steps, the algorithm changes nature, and passes from an inertial algorithm to a steepest descent method. This combined with the Hessian-driven damping makes it possible to considerably reduce the

oscillations: one benefits from the inertial effect at the beginning, then one passes to a method of gradient. In many ways, this closed loop control of the algorithm/dynamic has similarities to restart methods. Most importantly, the algorithm tolerates errors that are only supposed to converge to zero. It is a well known fact that there is a trade-off between fast convergence of optimization methods and their robustness to perturbations. Thus the algorithm is an interesting balance between fast convergence and robustness. This makes the algorithm a promising tool for dealing with stochastic/noisy situations in nonconvex, nonsmooth optimization. In addition, the technique that is developed is quite flexible. By relying on the threshold effect attached to dry damping, one can imagine controlling the dynamics, and thus switching to different regimes. The nonsmooth and nonconvex d.c. problem is also considered. Several questions require additional investigations, concerning for example general composite optimization problems, as well as the study of the associated stochastic algorithms. This is beyond the scope of this manuscript and will be the subject of further work.

References

[1] S. ADLY, H. ATTOUCH, *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping*, SIAM J. Optim., 30 (3) (2020), pp. 2134–2162.

[2] S. ADLY, H. ATTOUCH, *Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping*, J. Conv. Anal., 28 (2) (2021), pp. 281–310

[3] S. ADLY, H. ATTOUCH, *First-order inertial algorithms involving dry friction damping*, Math. Program, (2021) https://doi.org/10.1007/s10107-020-01613-y

[4] S. ADLY, H. ATTOUCH, 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, H. ATTOUCH, J. BOLTE, P. REDONT, *A second-order gradient-like dissipative dynamical system with Hessian-driven damping*, J. Math. Pures Appl., 81 (2002), pp. 747–779.

[6] 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.

[7] H. ATTOUCH, J. BOLTE, P. REDONT, A. SOUBEYRAN, *Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Lojasiewicz inequality*, Mathematics of Operations Research, 35 (2) (2010), pp. 438–457.

[8] H. ATTOUCH, J. BOLTE, B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, regularized Gauss-Seidel methods*, Math. Program., 137(1) (2013), pp. 91–129.

[9] H. ATTOUCH, R.I. BOŢ, E.R. CSETNEK, *Fast optimization via inertial dynamics with closed-loop damping*, Journal of the European Mathematical Society (JEMS), 2021, preprint available at hal-02910307.

[10] H. ATTOUCH, G. BUTTAZZO, G. MICHAILLE, *Variational Analysis in Sobolev and BV Spaces*: Applications to PDEs and Optimization, 2nd ed., MOS/SIAM Ser. Optim. 17, SIAM, Philadelphia, 2014.

[11] H. ATTOUCH, A. CABOT, *Convergence rates of inertial forward-backward algorithms*, SIAM J. Optim., 28 (1) (2018), pp. 849–874.

[12] H. ATTOUCH, A. CABOT, Z. CHBANI, H. RIAHI, *Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization*, JOTA, 179 (1) (2018), pp. 1–36 .

[13] H. ATTOUCH, Z. CHBANI, J. FADILI, H. RIAHI, *First-order optimization algorithms via inertial systems with Hessian driven damping*, Math. Program., (2020), https://doi.org/10.1007/s10107-020-01591-1.

[14] H. ATTOUCH, Z. CHBANI, J. PEYPOUQUET, 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, J. FADILI, V. KUNGURTSEV, *On the effect of perturbations, errors in first-order optimization methods with inertia and Hessian driven damping*, arXiv:2106.16159v1 [math.OC] 30 Jun 2021.

[16] H. ATTOUCH, J. PEYPOUQUET, P. REDONT, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differ. Equ., 261 (10), (2016), pp. 5734–5783.

[17] J.-F. AUJOL, CH. DOSSAL, *Stability of over-relaxations for the ForwardBackward algorithm, application to FISTA*, SIAM J. Optim., 25 (2015), pp. 24082433.

[18] J.-F. AUJOL, CH. DOSSAL, G. FORT, E. MOULINES, *Rates of Convergence of Perturbed FISTA-based algorithms*. 2019. hal-02182949. https://hal.archives-ouvertes.fr/hal-02182949

[19] J.-F. AUJOL, CH. DOSSAL, A. RONDEPIERRE, *Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization*. 2021. hal-02545245v2. https://hal.archives-ouvertes.fr/hal-02545245v2

[20] F. BACH, *Statistical machine learning and convex optimization*, StatMathAppli 2017, Fréjus - September 2017.

[21] M. BALTI, R. MAY, *Asymptotic for the perturbed heavy ball system with vanishing damping term*, Evol. Equ. Control Theory, 6 (2017), pp. 177–186.

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

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

[24] Y. BELLO-CRUZ, M. L. N. GONALVES, N. KRISLOCK, *On inexact accelerated proximal gradient methods with relative error rules*, preprint arXiv:2005.03766, (2020).

[25] J. BOLTE, A. DANIILIDIS, O. LEY, L. MAZET, *Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (6) (2010), pp. 3319-3363.

[26] R.I. BOŢ, E. R. CSETNEK, *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM J. Control Optim., 54 (3) (2016), pp. 1423-1443.

[27] R.I. BOŢ, E. R. CSETNEK, S.C. LÁSZLÓ, *A second order dynamical approach with variable damping to nonconvex smooth minimization*, to appear in Applicable Analysis, (2018).

[28] R.I. BOŢ, E. R. CSETNEK, S.C. LASZLÓ, *Tikhonov regularization of a second order dynamical system with Hessian damping*, Math. Program. (2020), https://doi.org/10.1007/s10107-020-01528-8.

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

[30] C. CASTERA, J. BOLTE, C. FÉVOTTE, E. PAUWELS, *An inertial Newton algorithm for deep learning*, (2019), preprint available at https://hal.inria.fr/hal-02140748/.

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

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

[33] L. VAN DEN DRIES, *Tame Topology and o-Minimal Structures*, London Mathematical Society, Lecture Note Series, Vol. 248., Cambridge University Press, Cambridge, UK, (1998).

[34] A. HARAUX, M. GHISI, M. GAMBINO, *Local and global smoothing effects for some linear hyperbolic equations with a strong dissipation*, Trans. Amer. Math. Soc. 368 (2016), 2039-2079.

[35] J.-B. HIRIAT-URRUTY, *How to Regularize a Difference of Convex Functions*, J. Math. Anal. Appl., 162 (1991), pp. 196–209.

[36] A.IOFFE, *An invitation to tame optimization*, SIAM J. Optim., 19(4) (2009), pp. 1894–1917.

[37] D. KIM, *Accelerated proximal point method for maximally monotone operators*, Math. Program., (2021).

[38] H.A. LE THI AND T. PHAM DINH, *The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems*, Ann. Oper. Res., 133 (2005), pp. 23-48.

[39] T. LIN, M.I. JORDAN, *A control-theoretic perspective on optimal high-order optimization*, (2019), arXiv:1912.07168v1.

[40] 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.

[41] Y. NESTEROV, *Introductory Lectures on Convex Optimization:* Appl. Optim. 87, Kluwer, Boston, MA, 2004.

[42] J. PEYPOUQUET, S. SORIN, *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, J. Convex Anal, 17 (3–4) (2010), pp. 1113–1163.

[43] T. PHAM DINH, E.B. SOUAD, *Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients*. J.B. Hiriart-Urruty (ed.) Fermat Days 85: Mathematics for Optimization, North-Holland Math. Stud., 129 (1986), pp. 249–271.

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

[45] M. SCHMIDT, N. LE ROUX, F. BACH, *Convergence rates of inexact proximal-gradient methods for convex optimization*. In: NIPS1125, (2011), Granada. HAL inria-00618152v3.

[46] B. SHI, S. S. DU, M. I. JORDAN, W. J. SU, *Understanding the acceleration phenomenon via high-resolution differential equations*, (2018), arXiv:1810.08907 [math.OC].

[47] M.V. SOLODOV, S.K. ZAVRIEV, *Error stability properties of generalized gradient-type algorithms*, J. Optim. Theory Appl., 98 (1998), pp. 663-680.

[48] W. SU, S. BOYD, E. J. CANDÈS, *A differential equation for modeling Nesterovs accelerated gradient method*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.

[49] J. TOLAND, *Duality in nonconvex optimization*, J. Math. Anal. Appl., 66 (1978), pp. 399–415.

[50] S. VILLA, S. SALZO, L. BALDASSARRES, A. VERRI, *Accelerated and inexact forwardbackward*, SIAM J. Optim., 23 (2013), pp. 1607-1633.