Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent

Hedy Attouch
IMAG, Univ. Montpellier, CNRS, Montpellier, France, hedy.attouch@umontpellier.fr

Radu Ioan Boţ
Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, radu.bot@univie.ac.at

Dang-Khoa Nguyen
Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City 700000, Vietnam, ndkhoa@hcmus.edu.vn
Vietnam National University, Ho Chi Minh City 700000, Vietnam

Abstract. In a Hilbert setting, we develop a gradient-based dynamic approach for fast solving convex optimization problems. By applying time scaling, averaging, and perturbation techniques to the continuous steepest descent (SD), we obtain high-resolution ODEs of the Nesterov and Ravine methods. These dynamics involve asymptotically vanishing viscous damping and Hessian driven damping (either in explicit or implicit form). Mathematical analysis does not require developing a Lyapunov analysis for inertial systems. We simply exploit classical convergence results for (SD) and its external perturbation version, then use tools of differential and integral calculus, including Jensen's inequality. The method is flexible and by way of illustration we show how it applies starting from other important dynamics in optimization. We consider the case where the initial dynamics is the regularized Newton method, then the case where the starting dynamics is the differential inclusion associated with a convex lower semi-continuous potential, and finally we show that the technique can be naturally extended to the case of a monotone cocoercive operator. Our approach leads to parallel algorithmic results, which we study in the case of fast gradient and proximal algorithms. Our averaging technique shows new links between the Nesterov and Ravine methods.

Key words: Fast convex optimization, Damped inertial dynamic, Time scaling, Averaging, Nesterov and Ravine algorithms, Hessian driven damping, Proximal algorithms

MSC2000 subject classification: Primary: 37N40, 90C25; secondary: 37M15, 46N10, 65K10

OR/MS subject classification: Primary: programming: nonlinear: convex; secondary: programming: nonlinear: algorithms

1. Introduction. In a real Hilbert space $\mathcal{H}$, we develop a general dynamic approach whose objective is the rapid resolution of convex optimization problems via first-order methods. We consider the problem

$$\min\{f(x) : x \in \mathcal{H}\}, \tag{1}$$

where, throughout the paper, we make the following assumptions on the function f to be minimized

$$(\mathcal{A}) \begin{cases} f : \mathcal{H} \to \mathbb{R} \text{ is a convex function of class } C^1;\ S = \arg\min_{\mathcal{H}} f \neq \emptyset; \\ \nabla f \text{ is Lipschitz continuous on the bounded sets of } \mathcal{H}. \end{cases} \tag{2}$$

Our study is part of the close links between dissipative dynamical systems and optimization algorithms, the latter being obtained by temporal discretization of the continuous dynamics. The dynamic system approach makes it possible to harmoniously combine the ingredients of our approach, namely time scaling, averaging, and perturbation techniques to finally obtain inertial dynamics with proven fast optimization properties.

Let us recall the mainstream for the study of accelerated gradient methods. Then, in the next section, we will describe our approach, which is novel in many aspects. The classical approach can be traced back to Gelfand and Tsetlin [37], then Polyak [43, 44], and Attouch, Goudou and Redont [18], where fast first-order optimization algorithms are naturally linked to damped inertial dynamics via temporal discretization. In this approach, which is very inspired by mechanics, PDE's, and control theory the focus is on the design of the damping term of the inertial dynamic, and its asymptotic stabilization effect.

In recent years, for the minimization of general convex differentiable functions, an in-depth study has been carried out linking the accelerated gradient method of Nesterov [40, 41] to inertial dynamics with vanishing viscosity (the viscous damping coefficient tends to zero as $t \to +\infty$). That is the Su-Boyd-Candés dynamical system [48]

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \tag{3}$$

where fast optimization is obtained by taking $\alpha \geq 3$. In addition to viscous damping, taking into account geometric damping (in our situation it is driven by the Hessian of the function to be minimized) makes it possible to improve the performance of these methods by attenuating the oscillations. That is the dynamic

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta \nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0, \tag{4}$$

considered by Attouch, Peypouquet, and Redont in [22]. The central role played by these dynamics has been confirmed by their recent interpretation as the low and high resolution of the Nesterov and Ravine accelerated gradient methods, see Shi, Du, Jordan, and Su [47], Attouch, Chbani, Fadili, and Riahi [9, 10]. The study of the fast convergence properties of these dynamics and algorithms is based on Lyapunov analysis. But this approach does not give a simple explanation of the choice of dynamics and algorithms. Moreover the Lyapunov analysis is not trivial, and there is no general rule to find the Lyapunov functions. Still, these techniques have allowed a deep understanding of dynamics and accelerated algorithms for optimization, see [3, 4, 6, 7, 9, 11, 12, 16, 20, 28, 30, 33–35, 47, 48] for some recent contributions.

2. Time scaling and averaging approach.

We propose a new approach to these questions which is based on a combination of time scaling, averaging, and perturbation techniques. The basis property which underlies our approach is the fact that by time scaling one can always accelerate a dynamic. The kinematic interpretation is clear. By changing the time clock the trajectories are travelled more or less quickly. The averaging technique shifts from a dynamic of order n to a dynamic of the immediately higher order $n + 1$. While preserving the asymptotic convergence properties, this makes it possible to achieve a dynamic where the coefficient in front of the gradient is fixed, and therefore suitable for developing an accelerated gradient algorithm. The perturbation technique gives flexibility to these methods, exploiting the fact that the convergence properties are preserved when adding an external perturbation which vanishes asymptotically sufficiently fast.

As the basic starting dynamic used in this paper, we consider the classical continuous steepest descent

$$\text{(SD)} \quad \dot{z}(t) + \nabla f(z(t)) = 0. \tag{5}$$

Under the standing assumption $(\mathcal{A})$ on f we know that, for any $z_0 \in \mathcal{H}$, there exists a unique classical global solution $z \in C^1([t_0, +\infty[; \mathcal{H})$ of (SD) satisfying $z(t_0) = z_0$, see [5, Theorem 17.1.1].

Throughout the paper we fix $t_0 > 0$ as the origin of time due to the singularity at the origin of $\frac{\alpha}{t}$, which occurs in the formulation and the analysis of the second-order dynamical systems (see, for instance, (3) and (4)).

2.1. Classical facts concerning the continuous steepest descent.

We will have to consider the gradient system with an external perturbation. The following theorem investigates its asymptotic properties under appropriate conditions on the perturbation term. The proof is provided in the Appendix.

4

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

THEOREM 1. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ satisfies $(\mathcal{A})$. Let $z : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system*

$$\dot{z}(t) + \nabla f(z(t)) = g(t), \tag{6}$$

where $g : [t_0, +\infty[\to \mathcal{H}$ is such that

$$\int_{t_0}^{+\infty} \|g(t)\| \, dt < +\infty \text{ and } \int_{t_0}^{+\infty} t \|g(t)\|^2 \, dt < +\infty. \tag{7}$$

Then the following statements are true:

(i) *(integral estimate of the velocities) $\int_{t_0}^{+\infty} t \|\dot{z}(t)\|^2 \, dt < +\infty$;*

(ii) *(integral estimate of the gradients) $\int_{t_0}^{+\infty} t \|\nabla f(z(t))\|^2 \, dt < +\infty$;*

(iii) *(integral estimate of the values) $\int_{t_0}^{+\infty} (f(z(t)) - \inf_{\mathcal{H}} f) \, dt < +\infty$;*

(iv) *(convergence of the values towards the minimal value) $f(z(t)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{t}\right)$ as $t \to +\infty$;*

(v) *the solution trajectory $z(t)$ converges weakly as $t \to +\infty$, and its limit belongs to $S = \arg\min f$.*

If

$$\int_{t_0}^{+\infty} t^2 \|g(t)\|^2 \, dt < +\infty,$$

then

(vi) *(convergence of the gradients towards zero) $\|\nabla f(z(t))\| = o\left(\frac{1}{t}\right)$ as $t \to +\infty$.*

2.2. General time scaling and averaging.

Let us make the change of time variable $t = \tau(s)$ in the dynamic (SD),

$$\text{(SD)} \qquad \dot{z}(t) + \nabla f(z(t)) = 0 \tag{8}$$

where $\tau(\cdot)$ is an increasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$, which is continuously differentiable, and which satisfies $\lim_{s \to +\infty} \tau(s) = +\infty$. Set $y(s) := z(\tau(s))$ and s_0 be such that $t_0 = \tau(s_0)$. On the one hand, by the derivation chain rule, we have

$$\dot{y}(s) = \dot{\tau}(s)\dot{z}(\tau(s)). \tag{9}$$

On the other hand, setting $t = \tau(s)$ in (SD) gives

$$\dot{z}(\tau(s)) + \nabla f(z(\tau(s))) = 0. \tag{10}$$

According to (9) and (10), we obtain

$$\dot{y}(s) + \dot{\tau}(s)\nabla f(y(s)) = 0. \tag{11}$$

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

5

The convergence rate becomes (take $t = \tau(s)$ in (iv) of Theorem 1 with $g \equiv 0$)

$$f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{\tau(s)}\right) \text{ as } s \to +\infty. \tag{12}$$

By introducing a function $\tau(s)$ that grows faster than the identity (namely $\tau(s) \geq s$), we have accelerated the dynamic, passing from the asymptotic convergence rate $1/t$ for (SD) to $1/\tau(s)$ for (11). The price to pay is that we no longer have an autonomous dynamic in (11), with as major drawback the fact that the coefficient in front of the gradient term tends to infinity as $s \to +\infty$. This prevents from using gradient methods to discretize it. Recall that for gradient methods the step size has to be less than or equal to twice the inverse of the Lipschitz constant of the gradient. To overcome this we come with the second step of our method which is averaging.

Several recent papers have been devoted to the role of the averaging techniques in the acceleration of optimization algorithms. In [42] Nesterov designed an algorithm that makes use of an averaging step for the gradients of the generated sequence. In [46] Scieur, D'Aspremont, and Bach developed a universal acceleration method based on averaging via polynomial techniques. In [45] Poveda and Li developed another universal acceleration method via averaging technique for singularly perturbed hybrid dynamical systems with restarting mechanisms. Let us explain some simple and, as far as we know, new ideas concerning the averaging methods based on differential equations. The use of differential equations allows us to use differential and integral calculus, which turns out to be a flexible approach.

Let us attach to $y(\cdot)$ the new function $x : [s_0, +\infty[\to \mathcal{H}$ defined by

$$\dot{x}(s) + \frac{1}{\dot{\tau}(s)}(x(s) - y(s)) = 0, \tag{13}$$

with $x(s_0) = x_0$ given in $\mathcal{H}$.

This means that, given $y(\cdot)$, one can recover $x(\cdot)$ as an exponentially average of it as

$$x(s) = \frac{x_0 + \int_{s_0}^{s} \frac{y(u)}{\dot{\tau}(u)} \exp\left(\int_{s_0}^{u} \frac{dr}{\dot{\tau}(r)}\right) du}{\exp\left(\int_{s_0}^{s} \frac{du}{\dot{\tau}(u)}\right)}. \tag{14}$$

From

$$y(s) = x(s) + \dot{\tau}(s)\dot{x}(s), \tag{15}$$

by temporal derivation of (15) we get

$$\dot{y}(s) = \dot{x}(s) + \ddot{\tau}(s)\dot{x}(s) + \dot{\tau}(s)\ddot{x}(s). \tag{16}$$

6

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

Replacing $\dot{y}(s)$ as given by (16) in (11) we get

$$\dot{\tau}(s)\ddot{x}(s) + (1 + \ddot{\tau}(s))\dot{x}(s) + \dot{\tau}(s)\nabla f(y(s)) = 0. \tag{17}$$

After simplification we obtain

$$\ddot{x}(s) + \frac{1 + \ddot{\tau}(s)}{\dot{\tau}(s)}\dot{x}(s) + \nabla f\Big(x(s) + \dot{\tau}(s)\dot{x}(s)\Big) = 0. \tag{18}$$

In doing so, we passed from the first-order differential equation (11) to the second-order differential equation (18), with the advantage that now the coefficient in front of the gradient is fixed. Let us now particularize the time scale $\tau(\cdot)$.

Open loop control: link with Nesterov method. According to the Su, Boyd, and Candés [48] model for the Nesterov method we consider the case where the viscous damping coefficient in (18) satisfies

$$\frac{1 + \ddot{\tau}(s)}{\dot{\tau}(s)} = \frac{\alpha}{s}. \tag{19}$$

For the solution of the second-order linear differential equation we get

$$\tau(s) = \frac{s^2}{2(\alpha - 1)} + C_1 s^{\alpha+1} + C_2, \tag{20}$$

for C_1, C_2 real constants. This leads to the choice $\alpha > 1$. To achieve orders of convergence independent of the parameter α, we set $C_1 = C_2 := 0$. This gives

$$\tau(s) = \frac{s^2}{2(\alpha - 1)}. \tag{21}$$

Equation (11) becomes

$$\dot{y}(s) + \frac{s}{\alpha - 1}\nabla f(y(s)) = 0, \tag{22}$$

and (12) gives

$$f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right) \text{ as } s \to +\infty. \tag{23}$$

From (18) we obtain

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \frac{s}{\alpha - 1}\dot{x}(s)\right) = 0. \tag{24}$$

In doing so, we passed from the first-order differential equation

$$\dot{y}(s) + \frac{s}{\alpha - 1}\nabla f(y(s)) = 0 \tag{25}$$

to the second-order differential equation (24), with the advantage that now the coefficient in front of the gradient is fixed. Of course, we have to prove that the fast convergence rates are preserved.

The above dynamic (24) is a particular case of the Inertial System with Implicit Hessian Damping

$$\text{(ISIHD)} \quad \ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \beta(s)\dot{x}(s)\right) = 0, \tag{26}$$

considered by Alecsa, László, and Pinţa in [2], see also Attouch, Fadili, and Kungurtsev [17] in the perturbed case. The rationale justifying the use of the term "implicit" comes from the observation that by Taylor expansion (as $s \to +\infty$ we have $\dot{x}(s) \to 0$) one has

$$\nabla f\left(x(s) + \beta(s)\dot{x}(s)\right) \approx \nabla f(x(s)) + \beta(s)\nabla^2 f(x(s))\dot{x}(s),$$

hence making the Hessian damping appear indirectly. We can therefore expect the dynamic (24) to have an asymptotic behavior similar to that of

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \frac{s}{\alpha - 1}\nabla^2 f(x(s))\dot{x}(s) + \nabla f(x(s)) = 0. \tag{27}$$

Closed loop control. The continuous steepest descent system (SD) provides several quantities which are strictly increasing and converge to $+\infty$, so they are eligible for time scaling. Since $\|\nabla f(z(t))\|$ is monotonically decreasing to zero, taking

$$t = \tau(s) = \frac{1}{\|\nabla f(z(s))\|^p},$$

is eligible for any $p > 0$. According to (SD) we have that the same holds true for

$$t = \tau(s) = \frac{1}{\|\dot{z}(s)\|^p}.$$

Since τ comes into the scaling technique via its derivative, we can also consider for $p > 0$

$$t = \dot{\tau}(s) = \frac{1}{\|\dot{z}(s)\|^p}.$$

Alternatively, we may also consider building τ with the help of the function values since $f(z(\cdot))$ is decreasing.

All these choices will lead to closed-loop dynamical systems which have attracted of many researchers recently, see for instance [4, 38].

8

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

2.3. Convergence rates. The above results suggest that similar results are valid for (18). We will examine the convergence rates satisfied by the trajectories of (18) simply by using time scale and averaging arguments. We do not make any other Lyapunov analysis, we just use the Lyapunov analysis of the steepest descent, which was recalled in Subsection 2.1. Let us state our first result which concerns the implicit Hessian driven damping dynamics.

THEOREM 2. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ satisfies ($\mathcal{A}$). Let $x : [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system*

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \frac{s}{\alpha-1}\dot{x}(s)\right) = 0. \tag{28}$$

Assume that $\alpha > 1$. Then the following statements are true:

(i) *(integral estimate of the gradients)* $\quad \int_{s_0}^{+\infty} s^3 \left\| \nabla f\left(x(s) + \frac{s}{\alpha-1}\dot{x}(s)\right) \right\|^2 ds < +\infty;$

(ii) *(convergence of the gradients towards zero)* $\quad \left\| \nabla f\left(x(s) + \frac{s}{\alpha-1}\dot{x}(s)\right) \right\| = o\left(\frac{1}{s^2}\right)$ *as $s \to +\infty$;*

(iii) *(convergence of the velocities towards zero)* $\quad \|\dot{x}(s)\| = o\left(\frac{1}{s}\right)$ *as $s \to +\infty$;*

(iv) *the solution trajectory $x(s)$ converges weakly as $s \to +\infty$, and its limit belongs to $S = \arg\min f$.*

   If $\alpha > 3$, then

(v) *(convergence of the values towards the minimal value)* $\quad f(x(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right)$ *as $s \to +\infty$.*

   F irst observe that the time scaling and averaging operations allow us to obtain any solution trajectory of (28). To this end we need to choose adequately the initial data in the first order evolution equations (25) and (13) which are respectively attached to the scaling and averaging operations. So let us give a solution trajectory $x(\cdot)$ of (28) which satisfies the Cauchy data $x(s_0) = x_0$ and $\dot{x}(s_0) = x_1$. Let us verify that this solution is reached by taking successively

$$\begin{cases} \dot{y}(s) + \frac{s}{\alpha-1}\nabla f(y(s)) = 0 \\ y(s_0) = x_0 + \frac{s_0}{\alpha-1}x_1, \end{cases} \tag{29}$$

then

$$\begin{cases} \dot{x}(s) + \frac{\alpha-1}{s}(x(s) - y(s)) = 0 \\ x(s_0) = x_0. \end{cases} \tag{30}$$

Indeed, (30) gives $\dot{x}(s_0) = \frac{\alpha-1}{s_0}(y(s_0) - x(s_0))$, which, by the second equation of (29), leads to $\dot{x}(s_0) = x_1$.

Let us first interpret the passage from y to x as an averaging process. To this end rewrite (30) as

$$s\dot{x}(s) + (\alpha - 1)x(s) = (\alpha - 1)y(s). \tag{31}$$

After multiplication of (31) by $s^{\alpha - 2}$, we get equivalently

$$s^{\alpha - 1}\dot{x}(s) + (\alpha - 1)s^{\alpha - 2}x(s) = (\alpha - 1)s^{\alpha - 2}y(s), \tag{32}$$

that is

$$\frac{d}{ds}\left(s^{\alpha - 1}x(s)\right) = (\alpha - 1)s^{\alpha - 2}y(s). \tag{33}$$

By integrating (33) from s_0 to s, and according to $x(s_0) = x_0$, we obtain

$$x(s) = \frac{s_0^{\alpha - 1}}{s^{\alpha - 1}}x_0 + \frac{\alpha - 1}{s^{\alpha - 1}}\int_{s_0}^{s} u^{\alpha - 2}y(u)\,du \tag{34}$$

$$= \frac{s_0^{\alpha - 1}}{s^{\alpha - 1}}y(s_0) + \frac{\alpha - 1}{s^{\alpha - 1}}\int_{s_0}^{s} u^{\alpha - 2}y(u)\,du - \frac{s_0^{\alpha}}{(\alpha - 1)s^{\alpha - 1}}x_1 \tag{35}$$

where the last equality follows from the choice of $y(s_0)$ as given by (29). Then, observe that $x(s)$ can be simply written as follows

$$x(s) = \int_{s_0}^{s} y(u)\,d\mu_s(u) + \xi(s), \tag{36}$$

where μ_s is the positive Radon measure on $[s_0, s]$ defined by

$$\mu_s = \frac{s_0^{\alpha - 1}}{s^{\alpha - 1}}\delta_{s_0} + (\alpha - 1)\frac{u^{\alpha - 2}}{s^{\alpha - 1}}\,du,$$

where δ_{s_0} is the Dirac measure at s_0, and

$$\xi(s) := -\frac{s_0^{\alpha}}{(\alpha - 1)s^{\alpha - 1}}x_1$$

is considered as a small perturbation for large values of s.

(i) & (ii) According to the energy estimates for the continuous steepest descent system, see Theorem 1, we have that for any solution trajectory of

$$\text{(SD)} \quad \dot{z}(t) + \nabla f(z(t)) = 0 \tag{37}$$

it holds

$$\int_{t_0}^{+\infty} t \,\|\dot{z}(t)\|^2 \, dt < +\infty, \quad \int_{t_0}^{+\infty} t \|\nabla f(z(t))\|^2 dt < +\infty, \text{ and } \|\nabla f(z(t))\| = o\left(\frac{1}{t}\right) \text{ as } t \to +\infty. \quad (38)$$

Notice that from the unperturbed (SD) we also have

$$\|\dot{z}(t)\| = o\left(\frac{1}{t}\right) \text{ as } t \to +\infty. \quad (39)$$

After making the change of time variable

$$t = \tau(s) = \frac{s^2}{2(\alpha - 1)},$$

and setting $z(\tau(s)) = y(s)$, the second integral estimate above becomes

$$\int_{s_0}^{+\infty} \tau(s)\|\nabla f(z(\tau(s))\|^2 \dot{\tau}(s)ds < +\infty, \quad (40)$$

that is

$$\int_{s_0}^{+\infty} s^3 \|\nabla f(y(s))\|^2 ds < +\infty. \quad (41)$$

Replacing y by its equivalent formulation $y(s) = x(s) + \frac{s}{\alpha-1}\dot{x}(s)$ in (38) and (41) gives the claims.

(iii) From (39) we have that there exists a positive function $\varepsilon_0(\cdot)$ which satisfies $\lim_{s \to +\infty} \varepsilon_0(s) = 0$ such that

$$\|\dot{z}(\tau(s))\| = \frac{\varepsilon_0(\tau(s))}{\tau(s)}, \quad (42)$$

and since $\dot{y}(s) = \dot{\tau}(s)\dot{z}(\tau(s))$ we get

$$\|\dot{y}(s)\| = \frac{\varepsilon_0(\tau(s))\dot{\tau}(s)}{\tau(s)}. \quad (43)$$

From $\tau(s) = \frac{s^2}{2(\alpha-1)}$, we get

$$\|\dot{y}(s)\| = \frac{\varepsilon_1(s)}{s}, \quad (44)$$

with $\varepsilon_1(s) = 2\varepsilon_0(\tau(s))$ which tends to zero as $s \to +\infty$. Let us now establish a similar estimate for $\|\dot{x}(s)\|$. According to (34)

$$x(s) = \frac{s_0^{\alpha-1}}{s^{\alpha-1}}x_0 + \frac{\alpha-1}{s^{\alpha-1}} \int_{s_0}^{s} u^{\alpha-2} y(u)du. \quad (45)$$

Let us derivate this expression. We get

$$\dot{x}(s) = -\frac{(\alpha-1)s_0^{\alpha-1}}{s^\alpha}x_0 - \frac{(\alpha-1)^2}{s^\alpha} \int_{s_0}^{s} u^{\alpha-2} y(u)du + (\alpha-1)\frac{1}{s}y(s). \quad (46)$$

Let us reformulate this expression in terms of $\dot{y}(s)$, which is the quantity whose speed of convergence is known by (44). Indeed, by integration by parts we have

$$\int_{s_0}^{s} u^{\alpha-1}\dot{y}(u)du = s^{\alpha-1}y(s) - s_0^{\alpha-1}y(s_0) - (\alpha-1)\int_{s_0}^{s} u^{\alpha-2}y(u)du.$$

After multiplication of the above expression by $\frac{\alpha-1}{s^{\alpha}}$, and according to $y(s_0) = x_0 + \frac{s_0}{\alpha-1}x_1$, we get

$$\frac{\alpha-1}{s^{\alpha}}\int_{s_0}^{s} u^{\alpha-1}\dot{y}(u)du = (\alpha-1)\frac{1}{s}y(s) - \frac{(\alpha-1)s_0^{\alpha-1}}{s^{\alpha}}x_0 - \frac{s_0^{\alpha}}{s^{\alpha}}x_1 - \frac{(\alpha-1)^2}{s^{\alpha}}\int_{s_0}^{s} u^{\alpha-2}y(u)du. \quad (47)$$

Comparing (46) and (47) we get

$$\dot{x}(s) = \frac{\alpha-1}{s^{\alpha}}\int_{s_0}^{s} u^{\alpha-1}\dot{y}(u)du + \frac{s_0^{\alpha}}{s^{\alpha}}x_1.$$

According to (44), we obtain successively

$$\|\dot{x}(s)\| \leq \frac{\alpha-1}{s^{\alpha}}\int_{s_0}^{s} u^{\alpha-1}\|\dot{y}(u)\|\,du + \frac{s_0^{\alpha}}{s^{\alpha}}\|x_1\| = \frac{\alpha-1}{s^{\alpha}}\int_{s_0}^{s} u^{\alpha-2}\varepsilon_1(u)du + \frac{s_0^{\alpha}}{s^{\alpha}}\|x_1\|. \quad (48)$$

After multiplication of (48) by s, we obtain

$$s\|\dot{x}(s)\| \leq \frac{\alpha-1}{s^{\alpha-1}}\int_{s_0}^{s} u^{\alpha-2}\varepsilon_1(u)du + \frac{s_0^{\alpha}}{s^{\alpha-1}}\|x_1\|,$$

and Lemma 4 gives us $\lim_{s\to+\infty}\|s\dot{x}(s)\| = 0$.

(iv) For trajectory convergence, following the approach of the paper, we do not perform a Lyapunov analysis, but take advantage of the fact that the trajectory $z(s)$ of the steepest descent dynamics converges weakly towards a solution $x_* \in S$ as $s \to +\infty$. This immediately implies that $y(s) = z(u(s))$ converges weakly to x_* as $s \to +\infty$. In other words, for each $v \in \mathcal{H}$

$$\langle y(s), v \rangle \to \langle x_*, v \rangle \text{ as } s \to +\infty.$$

To pass from the convergence of y to that of x, we use the interpretation of x as an average of y plus a negligible term. The convergence then results from the general property which says that convergence entails ergodic convergence. Let us make this precise. Since the perturbation term ξ is negligible as $s \to +\infty$, we have by definition of $x(s)$

$$x(s) \sim \frac{s_0^{\alpha-1}}{s^{\alpha-1}}y(s_0) + \frac{\alpha-1}{s^{\alpha-1}}\int_{s_0}^{s} u^{\alpha-2}y(u)du.$$

12

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

After elementary calculus, we just need to prove that for $a(\cdot)$ be a positive real valued function which verifies $\lim_{s \to +\infty} a(s) = 0$, then $\lim_{s \to +\infty} A(s) = 0$, where

$$A(s) = \frac{1}{s^{\alpha-1}} \int_{s_0}^{s} u^{\alpha-2} a(u) du,$$

which is true according to Lemma 4.

(v) First, let us get rid of the perturbation term. From (iv) we deduce that $x(\cdot)$ is also bounded. Since ∇f is Lipschitz continuous on the bounded sets, it follows that ∇f is bounded on the bounded sets, which immediately implies that f is also Lipschitz continuous on the bounded sets. Therefore there exists $L_f > 0$ such that

$$f(x(s)) - f\left(\int_{s_0}^{s} y(u) \, d\mu_s(u)\right) \leq L_f \|\xi(s)\| \leq \frac{C_f}{s^{\alpha-1}} \quad \forall s \geq s_0,$$

for $C_f := \frac{L_f s_0^{\alpha}}{\alpha-1} \|x_1\|$, which leads to

$$f(x(s)) - \inf_{\mathcal{H}} f \leq f\left(\int_{s_0}^{s} y(u) \, d\mu_s(u)\right) - \inf_{\mathcal{H}} f + \frac{C_f}{s^{\alpha-1}} \quad \forall s \geq s_0. \tag{49}$$

We are therefore reduced to examining the convergence rate of $f\left(\int_{s_0}^{s} y(u) d\mu_s(u)\right) - \inf f$ towards zero as $s \to +\infty$. We have that μ_s is a positive Radon measure on $[s_0, s]$ whose total mass is equal to 1. It is therefore a probability measure, and $\int_{s_0}^{s} y(u) \, d\mu_s(u)$ is obtained by **averaging** the trajectory $y(\cdot)$ on $[s_0, s]$ with respect to μ_s. From there, we can deduce fast convergence properties for the solution trajectories of (28). According to the convexity of f, and Jensen's inequality, we obtain that

$$f\left(\int_{s_0}^{s} y(u) \, d\mu_s(u)\right) - \inf_{\mathcal{H}} f \leq \int_{s_0}^{s} (f(y(u)) - \inf_{\mathcal{H}} f) \, d\mu_s(u). \tag{50}$$

It follows from (23) that there exists a positive function $\varepsilon(\cdot)$ which satisfies $\lim_{s \to +\infty} \varepsilon(s) = 0$ such that

$$f(y(s)) - \inf_{\mathcal{H}} f = \frac{\varepsilon(s)}{s^2}. \tag{51}$$

According to the definition of μ_s, it yields for all $s \geq s_0$

$$f\left(\int_{s_0}^{s} y(u) \, d\mu_s(u)\right) - \inf_{\mathcal{H}} f \leq \int_{s_0}^{s} \frac{\varepsilon(u)}{u^2} d\mu_s(u) = \frac{s_0^{\alpha-3}}{s^{\alpha-1}} + \frac{\alpha-1}{s^{\alpha-1}} \int_{s_0}^{s} \varepsilon(u) u^{\alpha-4} du.$$

By making use of (49), we deduce

$$s^2 (f(x(s)) - \inf_{\mathcal{H}} f) \leq \frac{C_f + s_0^{\alpha-3}}{s^{\alpha-3}} + \frac{\alpha-1}{s^{\alpha-3}} \int_{s_0}^{s} \varepsilon(u) u^{\alpha-4} du.$$

Therefore, for $\alpha > 3$ we get

$$\limsup_{s \to +\infty} s^2 (f(x(s)) - \inf_{\mathcal{H}} f) \leq (\alpha - 1) \limsup_{s \to +\infty} \frac{1}{s^{\alpha-3}} \int_{s_0}^{s} \varepsilon(u) u^{\alpha-4} du.$$

Finally, we just need to apply Lemma 4 to prove the claim.   Q.E.D.

REMARK 1.   We emphasize the fact that the convergence of the solution trajectories of the damped inertial system (28) is valid under the condition $\alpha > 1$. This contrasts with the stronger condition $\alpha > 3$ which is required to obtain the convergence of the trajectories for the system without Hessian driven damping

$$\ddot{x}(s) + \frac{\alpha}{s} \dot{x}(s) + \nabla f(x(s)) = 0, \tag{52}$$

and which is the low-resolution ODE of Nesterov's accelerated gradient method. This property is related to the presence of the Hessian driven damping and the particular form of the corresponding coefficient $\frac{s}{\alpha-1}$. Moreover, it has been established in [11, 20] that the choice $\alpha > 3$ provides a convergence rate of the values of $o\left(1/s^2\right)$ as $s \to +\infty$, and therefore it improves the convergence rate of Nesterov's accelrated gradient method of $O\left(1/s^2\right)$ as $s \to +\infty$. As seen, the time scale and averaging approach is flexible enough to also provide the convergence rate $o\left(1/s^2\right)$ as $s \to +\infty$, in the case $\alpha > 3$.

REMARK 2.   We also emphasize that, taking $C_1 > 0$ in (20), for every given $\alpha > 1$, a second-order dynamical system of the form (18) with a damping coefficient of $\frac{\alpha}{s}$ and convergence rate for the function values of $o\left(1/s^{\alpha+1}\right)$, as $s \to +\infty$, can be defined. Its asymptotic properties can be derived from those of the first-order dynamical system (11) by using the representation (14) and techniques similar to those in the proof of Theorem 2. However, there would be no guarantee that argument $x(s) + \dot{\tau}(s)\dot{x}(s)$ of the gradient in (18) is bounded, which could lead to major instabilities. This is because, as for the velocities of $z(\cdot)$ and $y(\cdot)$, the convergence rate of $\dot{x}(s)$ is of $o(1/s)$ as $s \to +\infty$, while $\dot{\tau}(s) \sim s^{\alpha}$ for $\alpha > 1$.

REMARK 3.   The time scale and averaging technique provides a new approach which allows to transfer the convergence rate of the function value along the trajectory $y(s)$ obtained via the time scaling of (SD) to the function value along the trajectory $x(s)$ that can be seen as the average of the initial one. This phenomenon can be seen in analogy with the manner the Nesterov method and the Ravine method behave to each other.

14

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

Recall that the Nesterov'acceleration gradient [40, 41] reads

$$(\forall k \geq 1) \quad \begin{cases} y_k & := x_k + \alpha_k \left(x_k - x_{k-1} \right) \\ x_{k+1} & := y_k - \lambda \nabla f \left(y_k \right) \end{cases}, \tag{53}$$

where $x_0 = x_1 \in \mathcal{H}$, $0 < \lambda \leq \frac{1}{L}$ and

$$\alpha_k := \frac{t_k - 1}{t_{k+1}} \quad \text{with} \quad \begin{cases} t_1 := 1 \\ t_{k+1}^2 - t_{k+1} \leq t_k^2 \quad \forall k \geq 1 \end{cases}.$$

The so-called Ravine sequence $(y_k)_{k \geq 1}$ is the one generated by the Ravine method introduced by Gelfand and Tsetlin [16, 37]. The values of f along the Ravine sequence converge fast towards $\inf_{\mathcal{H}} f$. In analogy with the continuous time approach, the rate of convergence can be transferred to the values of f along $(x_k)_{k \geq 0}$, which represents average sequence of $(y_k)_{k \geq 1}$.

Indeed, by induction arguments, we have for every $k \geq 1$

$$x_{k+1} = \frac{1}{1 + \alpha_{k+1}} y_{k+1} + \frac{\alpha_{k+1}}{1 + \alpha_{k+1}} x_k = \frac{1}{1 + \alpha_{k+1}} y_{k+1} + \frac{\alpha_{k+1}}{1 + \alpha_{k+1}} \left(\frac{1}{1 + \alpha_k} y_k + \frac{\alpha_k}{1 + \alpha_k} x_{k-1} \right)$$

$$= \cdots = \sum_{i=1}^{k+1} \theta_{k+1,i} y_i,$$

where the nonnegative weights $\left(\theta_{k+1,i} \right)_{1 \leq i \leq k+1}$ defined by

$$\theta_{k+1,i} := \frac{1}{1 + \alpha_{k+1}} \prod_{j=1}^{k+1-i} \frac{\alpha_{k+2-j}}{1 + \alpha_{k+1-j}} \quad \forall 1 \leq i \leq k \quad \text{and} \quad \theta_{k+1,k+1} := \frac{1}{1 + \alpha_{k+1}}$$

fulfill $\sum_{i=1}^{k+1} \theta_{k+1,i} = 1$.

In the following we give a direct proof for the fast convergence rate of the Ravine method, which can be seen as an alternative approach to [16]. Let $z_* \in S$. We define for every $k \geq 2$ the discrete energy function

$$E_k := t_k^2 \left(f \left(y_{k-1} \right) - \inf_{\mathcal{H}} f - \lambda \left(1 - \frac{L\lambda}{2} \right) \| \nabla f \left(y_{k-1} \right) \|^2 \right)$$

$$+ \frac{1}{2\lambda} \| (t_k - 1) \left(x_k - x_{k-1} \right) + x_k - z_* \|^2 \geq 0.$$

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

15

Let $k \geq 2$. By the convexity of f, it yields

$$f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle \quad \text{and} \quad \inf_{\mathcal{H}} f \geq f(y_k) + \langle \nabla f(y_k), z_* - y_k \rangle.$$

By multiplying the first inequality by $t_{k+1}^2 - t_{k+1} > 0$ and the second one by $t_{k+1} > 0$ and summing the resulting inequalities, it yields

$$\begin{aligned}
&\left(t_{k+1}^2 - t_{k+1}\right)(f(x_k) - \inf_{\mathcal{H}} f) \\
&\geq t_{k+1}^2 (f(y_k) - \inf_{\mathcal{H}} f) - t_{k+1} \langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - z_* \rangle.
\end{aligned} \tag{54}$$

Since $t_{k+1} y_k = t_{k+1} x_k + (t_k - 1)(x_k - x_{k-1})$, by denoting

$$u_k := (t_{k+1} - 1)(y_k - x_k) + y_k - z_* = t_{k+1}(y_k - x_k) + x_k - z_* = (t_k - 1)(x_k - x_{k-1}) + x_k - z_*,$$

we have

$$-\lambda t_{k+1} \nabla f(y_k) = t_{k+1}(x_{k+1} - y_k) = t_{k+1}(x_{k+1} - x_k) - (t_k - 1)(x_k - x_{k-1}) = u_{k+1} - u_k.$$

Therefore

$$\begin{aligned}
&-t_{k+1} \langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - z_* \rangle \\
&= \frac{1}{\lambda} \langle u_{k+1} - u_k, u_k \rangle = \frac{1}{2\lambda}\left(\|u_{k+1}\|^2 - \|u_{k+1} - u_k\|^2 - \|u_k\|^2\right) \\
&= \frac{1}{2\lambda}\left(\|(t_{k+1} - 1)(x_{k+1} - x_k) + x_{k+1} - z_*\|^2 - \|(t_k - 1)(x_k - x_{k-1}) + x_k - z_*\|^2\right) \\
&\quad - \frac{1}{2}\lambda t_{k+1}^2 \|\nabla f(y_k)\|^2.
\end{aligned}$$

Replacing this expression into (54), after some rearrangements we obtain

$$E_{k+1} \le \frac{1}{2\lambda} \|(t_k - 1)(x_k - x_{k-1}) + x_k - z_*\|^2 + \left(t_{k+1}^2 - t_{k+1}\right)(f(x_k) - \inf_{\mathcal{H}} f)$$
$$- \frac{1}{2}\lambda(1 - L\lambda) t_{k+1}^2 \|\nabla f(y_k)\|^2.$$

In addition, the Descent Lemma gives

$$f(x_k) - \inf_{\mathcal{H}} f \le f(y_{k-1}) - \inf_{\mathcal{H}} f + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle + \frac{L}{2} \|x_k - y_{k-1}\|^2$$
$$= f(y_{k-1}) - \inf_{\mathcal{H}} f - \lambda\left(1 - \frac{L\lambda}{2}\right)\|\nabla f(y_{k-1})\|^2, \tag{55}$$

which further leads to the inequality

$$E_{k+1} \le E_k - \frac{1}{2}\lambda(1 - L\lambda) t_{k+1}^2 \|\nabla f(y_k)\|^2$$
$$+ \left(t_{k+1}^2 - t_{k+1} - t_k^2\right)\left(f(y_{k-1}) - \inf_{\mathcal{H}} f - \lambda\left(1 - \frac{L\lambda}{2}\right)\|\nabla f(y_{k-1})\|^2\right),$$

and it also proves that $E_k \ge 0$ for every $k \ge 2$. Therefore, when $0 < \lambda < \frac{1}{L}$, we can deduce

$$f(y_k) - \inf_{\mathcal{H}} f = O\left(\frac{1}{t_{k+1}^2}\right) = O\left(\frac{1}{(k+1)^2}\right) \text{ as } k \to +\infty,$$

which immediately transfers to

$$f(x_k) - \inf_{\mathcal{H}} f = O\left(\frac{1}{t_{k+1}^2}\right) = O\left(\frac{1}{(k+1)^2}\right) \text{ as } k \to +\infty.$$

3. Applications to various continuous dynamics. In this section we will show that the proposed approach which combines time scaling with averaging can be successfully applied beyond the classical steepest descent dynamical system.

3.1. Explicit Hessian damping. We consider an explicit Hessian driven damping version of the Su-Boyd-Candés dynamical system (3). As explained in (27) the following dynamic

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \frac{s}{\alpha+1}\frac{d}{ds}(\nabla f(y(s))) + \nabla f(y(s)) = 0 \tag{56}$$

comes naturally by performing a Taylor expansion in (28). Just note that to match the general writing of inertial dynamics for optimization, the above formula is obtained by taking $\alpha + 1$ instead of $\alpha - 1$.

The dynamic (56) is a particular case of the general dynamic on $[s_0, +\infty[$

$$(\text{DIN-AVD})_{\alpha,\beta,b} \quad \ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \beta(s)\nabla^2 f(y(s))\dot{y}(s) + b(s)\nabla f(y(s)) = 0,$$

studied by Attouch, Chbani, Fadili and Riahi in [9, 10]. Let us recall the convergence results obtained in [9, 10] by performing a Lyapunov analysis for $(\text{DIN} - \text{AVD})_{\alpha,\beta,b}$. They will serve as a comparison with our results.

THEOREM 3. *Assume that $\alpha \geq 1$. Set $w(s) := b(s) - \left(\dot{\beta}(s) + \frac{\beta(s)}{s}\right)$ and suppose that the following conditions are satisfied for every $s \geq s_0$*

$$(\mathcal{G}_2) \quad b(s) > \dot{\beta}(s) + \frac{\beta(s)}{s};$$

$$(\mathcal{G}_3) \quad s\dot{w}(s) \leq (\alpha - 3)w(s).$$

Then, for every solution trajectory $y : [s_0, +\infty[\to \mathcal{H}$ of $(\text{DIN} - \text{AVD})_{\alpha,\beta,b}$, the following statements are true:

(i) $f(y(s)) - \inf_{\mathcal{H}} f = O\left(\dfrac{1}{s^2 w(s)}\right)$ *as* $s \to +\infty$;

(ii) $\int_{s_0}^{+\infty} s^2 \beta(s) w(s) \|\nabla f(y(s))\|^2 ds < +\infty$;

(iii) $\int_{s_0}^{+\infty} s\left((\alpha - 3)w(s) - s\dot{w}(s)\right)(f(y(s)) - \inf_{\mathcal{H}} f) ds < +\infty$.

Let us specialize the above result to (28) by taking $b(s) = 1$ and $\beta(s) = \frac{s}{\alpha - 1}$ for every $s \geq s_0$. We get $w(\cdot) \equiv \frac{\alpha - 3}{\alpha - 1}$. Thus $(\mathcal{G}_2)$ is satisfied for $\alpha > 3$, while $(\mathcal{G}_3)$ reduces to $0 \leq \frac{(\alpha - 3)^2}{\alpha - 1}$, which is also satisfied. We thus get ffor the trajectories of (28)

$$f(y(s)) - \inf_{\mathcal{H}} f = O\left(\frac{1}{s^2}\right) \text{ as } s \to +\infty,$$

and

$$\int_{s_0}^{+\infty} s^3 \|\nabla f(y(s))\|^2 ds < +\infty \text{ and } \int_{s_0}^{+\infty} s(f(y(s)) - \inf_{\mathcal{H}} f) ds < +\infty.$$

In this particular setting we can deduce further (see [10, Theorem 2.1])

$$\int_{s_0}^{+\infty} s\|\dot{y}(s)\|^2 ds < +\infty; \quad f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right) \text{ and } \|\dot{y}(s)\| = o\left(\frac{1}{s}\right) \text{ as } s \to +\infty.$$

18

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

Let us show that this dynamic can be reached by time scaling of a perturbed version of the continuous steepest descent dynamical system. As for the implicit case, we make in the latter the change of time variable

$$\tau(s) = \frac{s^2}{2(\alpha+1)}.$$

Thus we obtain the rescaled steepest descent

$$\dot{y}(s) + \frac{s}{\alpha+1}\nabla f(y(s)) = 0, \tag{57}$$

with the convergence rate

$$f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right) \text{ as } s \to +\infty. \tag{58}$$

To obtain an explicit Hessian driven damping term, we take the derivative with respect to s in (57)

$$\ddot{y}(s) + \frac{1}{\alpha+1}\nabla f(y(s)) + \frac{s}{\alpha+1}\frac{d}{ds}(\nabla f(y(s))) = 0.$$

On the other hand, for every $s \geq \sigma_0$, by multiplying both sides of (57) by $\frac{\alpha}{s} > 0$, we get

$$\frac{\alpha}{s}\dot{y}(s) + \frac{\alpha}{\alpha+1}\nabla f(y(s)) = 0.$$

Summing up the two above equations above yields

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \frac{s}{\alpha+1}\frac{d}{ds}(\nabla f(y(s))) + \nabla f(y(s)) = 0,$$

which is precisely (56).

THEOREM 4.   *Suppose that $f: \mathcal{H} \to \mathbb{R}$ satisfies ($\mathcal{A}$). Let $y: [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system*

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \frac{s}{\alpha+1}\frac{d}{ds}(\nabla f(y(s))) + \nabla f(y(s)) = 0. \tag{59}$$

Assume that $\alpha > 1$. Then the following statements are true:

(i) *(integral estimate of the velocities)* $\int_{s_0}^{+\infty} s\,\|\dot{y}(s)\|^2\,ds < +\infty$;

(ii) *(integral estimate of the values)* $\int_{s_0}^{+\infty} s\,(f(y(s)) - \inf_{\mathcal{H}} f)\,ds < +\infty$;

(iii) *(integral estimate of the gradients)* $\int_{s_0}^{+\infty} s^3\,\|\nabla f(y(s))\|^2\,ds < +\infty$;

(iv) *(convergence of the values towards the minimal value)* $f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right)$ *as $s \to +\infty$;*

(v) *the solution trajectory $y(s)$ converges weakly as $s \to +\infty$, and its limit belongs to $S = \arg\min f$.*

   If $\alpha > 2$, then

(vi) *(convergence of the velocities towards zero)* $\quad \|\dot{y}(s)\| = o\left(\frac{1}{s}\right)$ *as* $s \to +\infty$;

(vii) *(convergence of the gradients towards zero)* $\quad \|\nabla f(y(s))\| = o\left(\frac{1}{s^2}\right)$ *as* $s \to +\infty$.

The proof consists in reversing the process described above which has permitted us to pass from (SD) to the damped inertial dynamic with explicit Hessian driven damping (59). Let y be an arbitrary solution trajectory of (59) which satisfies the Cauchy data $y(s_0) := y_0$ and $\dot{y}(s_0) := y_1$. Observe that

$$\frac{d}{ds}\left(s^\alpha \dot{y}(s)\right) = s^\alpha \ddot{y}(s) + \alpha s^{\alpha-1} \dot{y}(s)$$

$$\frac{d}{ds}\left(\frac{s^{\alpha+1}}{\alpha+1}\nabla f(y(s))\right) = \frac{s^{\alpha+1}}{\alpha+1}\frac{d}{ds}\nabla f(y(s)) + s^\alpha \nabla f(y(s)).$$

Hence, by multiplying both sides of (59) by $s^\alpha > 0$, we deduce that

$$\frac{d}{ds}\left(s^\alpha \dot{y}(s) + \frac{s^{\alpha+1}}{\alpha+1}\nabla f(y(s))\right) = 0.$$

This gives, equivalently

$$\dot{y}(s) + \frac{s}{\alpha+1}\nabla f(y(s)) = \frac{c_0}{s^\alpha}, \tag{60}$$

where $c_0 \in \mathcal{H}$ is a constant that can be determined from the Cauchy condition. Precisely, taking $s := s_0$ in (60) it yields

$$c_0 = s_0^\alpha y_1 + \frac{s_0^{\alpha+1}}{\alpha+1}\nabla f(y_0). \tag{61}$$

Next we show that time scaling makes it possible to link the solution trajectory $y(\cdot)$ of (59) to the solution trajectory of a perturbation of (SD). Set $z(t) = y(\sigma(t))$, where

$$\sigma(t) = \sqrt{2(\alpha+1)t}.$$

Therefore,

$$\dot{\sigma}(t) = \sqrt{\frac{\alpha+1}{2t}} \quad \text{and} \quad \dot{z}(t) = \dot{\sigma}(t)\dot{y}(\sigma(t)).$$

Moreover, notice that (60) can be equivalently written as

$$\frac{\alpha+1}{s}\dot{y}(s) + \nabla f(y(s)) = \frac{c_0(\alpha+1)}{s^{\alpha+1}}.$$

Setting $s = \sigma(t)$ gives with $c := c_0(\alpha+1) \in \mathcal{H}$

$$\frac{\alpha+1}{\sigma(t)}\frac{1}{\dot{\sigma}(t)}\dot{z}(t) + \nabla f(z(t)) = \frac{c}{\sigma(t)^{\alpha+1}}$$

20

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

or, equivalently,

$$\dot{z}(t) + \nabla f(z(t)) = \frac{c}{t^{\frac{\alpha+1}{2}}}.$$

This is nothing else than the perturbed continuous steepest descent system with the perturbation function $g \colon [t_0, +\infty[\to \mathcal{H}, g(t) = \frac{c}{t^{\frac{\alpha+1}{2}}}$, which obviously fulfills

$$\int_{t_0}^{+\infty} \|g(t)\| \, dt < +\infty \text{ and } \int_{t_0}^{+\infty} t \|g(t)\|^2 \, dt < +\infty,$$

for every $\alpha > 1$.

Furthermore,

$$\int_{t_0}^{+\infty} t^2 \|g(t)\|^2 \, dt = \int_{t_0}^{+\infty} \frac{\|c\|^2}{t^{\alpha-1}} \, dt < +\infty,$$

whenever $\alpha > 2$. All statements excepting (vi) follow from Theorem 1.

Going back to the perturbed continuous steepest descent (6), we see that

$$\|\dot{z}(t)\| \leq \|\nabla f(z(t))\| + \|g(t)\| = o\left(\frac{1}{t}\right) + o\left(\frac{1}{t^{\frac{\alpha+1}{2}}}\right) = o\left(\frac{1}{t}\right) \text{ as } t \to +\infty.$$

Taking $t = \tau(s) = \frac{s^2}{2(\alpha+1)}$, it yields

$$\|\dot{z}(\tau(s))\| = o\left(\frac{1}{s^2}\right) \text{ as } s \to +\infty,$$

which gives (vi), since $\dot{y}(s) = \dot{\tau}(s)\dot{z}(\tau(s)) = \frac{s}{\alpha+1}\dot{z}(\tau(s))$.   Q.E.D.

REMARK 4.   As in in Theorem 2, we see that the convergence of trajectory can be guaranteed for $\alpha > 1$, which is less restrictive than for the Su-Boyd-Candés system, the trajectory of which is known to convergence for $\alpha > 3$. This is another positive effect of the approach that combines time scaling and averaging.

3.2. Combining implicit and explicit Hessian driven damping.   As a starting dynamic we consider the regularized Newton dynamical system

$$\begin{cases} \lambda \dot{z}(t) + \dot{v}(t) + v(t) = 0 \\ v(t) = \nabla f(z(t)) \end{cases}. \tag{62}$$

It is a special case of the regularized Newton dynamic

$$\lambda(t)\dot{z}(t) + \nabla^2 f(z(t))\dot{z}(t) + \nabla f(z(t)) = 0. \tag{63}$$

This system has been studied by Attouch and Svaiter [24], Attouch, Redont, and Svaiter [23], Attouch, Marques Alves, and Svaiter [19] to solve monotone inclusions (that is, with a general maximally monotone operator A instead of ∇f). In this approach, the central question is the adjustment of the Levenberg-Marquardt regularization parameter $\lambda(\cdot)$ in front of the velocity term. Indeed, taking $\lim_{t \to +\infty} \lambda(t) = 0$ allows to be asymptotically close to the Newton method. Our situation concerns the simpler case $\lambda(t) \equiv \lambda > 0$ constant, which fits with the convergence properties proved in these papers, and the fact that this dynamic is well-posed. For precise statements concerning the asymptotic behavior of this system see [24, Theorem 4.1, Remark 4.2] and [24, Theorem 3.9]. For the sake of completeness, we provide below its convergence properties, that shows some improvements compared to the results in [24]. The proof of Theorem 5 is provided in the Appendix.

THEOREM 5. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ satisfies $(\mathcal{A})$. Let $z : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system* (62). *Then the following statements are true:*

(i) *(integral estimate of the values)* $\int_{t_0}^{+\infty} (f(z(t)) - \inf_{\mathcal{H}} f)\, dt < +\infty$;

(ii) *(convergence of the values towards the minimal value)* $f(z(t)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{t}\right)$ *as $t \to +\infty$;*

(iii) *(integral estimates of the velocities and the gradients)* $\int_{t_0}^{+\infty} t\, \|\dot{z}(t)\|^2\, dt < +\infty$ *and* $\int_{t_0}^{+\infty} t\, \|\dot{v}(t)\|^2\, dt < +\infty$, *which lead to* $\int_{t_0}^{+\infty} t\, \|v(t)\|^2\, dt = \int_{t_0}^{+\infty} t\, \|\nabla f(z(t))\|^2\, dt < +\infty$;

(iv) *(convergence of the velocities and the gradients towards zero)* $\|v(t)\| = \|\nabla f(z(t))\| = o\left(\frac{1}{t}\right)$ *as $t \to +\infty$, which lead to* $\|\dot{z}(t)\| = \|\dot{v}(t)\| = o\left(\frac{1}{t}\right)$ *as $t \to +\infty$;*

(v) *the solution trajectory $z(t)$ converges weakly as $t \to +\infty$, and its limit belongs to $S = \arg\min f$.*

REMARK 5. In [24, Theorem 4.1], the authors show that both $\dot{z}(\cdot)$ and $\dot{v}(\cdot)$, and consequently $v(\cdot)$, belong to $\mathbb{L}^2([t_0, +\infty[\, ;\mathcal{H})$. Since $t \mapsto \|v(t)\|$ is nonincreasing, in [24, Theorem 3.9] it is shown that $\|v(t)\| = o\left(\frac{1}{\sqrt{t}}\right)$ as $t \to +\infty$. By improving some integral estimates for the velocity and the gradient, we can provide a faster convergence rate, namely of $o\left(\frac{1}{t}\right)$ as $t \to +\infty$.

Let us make the change of time variable $t = \tau(s) = \frac{1}{2}\gamma s^2$ in the dynamic (62), where γ is a positive constant to be adjusted later. Set $y(s) := z(\tau(s))$, $w(s) := v(\tau(s)) = \nabla f(y(s))$, and $s_0 > 0$ be such that $t_0 = \tau(s_0)$. On the one hand, by the derivation chain rule, we have for every $s \geq s_0$

$$\dot{y}(s) = \dot{\tau}(s)\dot{z}(\tau(s)), \quad \dot{w}(s) = \dot{\tau}(s)\dot{v}(\tau(s)). \tag{64}$$

On the other hand, setting $t = \tau(s)$ in (62) gives

$$\begin{cases} \lambda\dot{z}(\tau(s)) + \dot{v}(\tau(s)) + v(\tau(s)) = 0 \\ v(\tau(s)) = \nabla f(z(\tau(s))). \end{cases} \tag{65}$$

22

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

According to (64), (65), and $\dot{\tau}(s) = \gamma s$, we obtain

$$
\begin{cases}
\frac{\lambda}{cs}\dot{y}(s) + \frac{1}{\gamma s}\dot{w}(s) + w(s) = 0 \\[2mm]
w(s) = \nabla f(y(s))
\end{cases}
\tag{66}
$$

or, equivalently,

$$
\begin{cases}
\lambda \dot{y}(s) + \dot{w}(s) + \gamma s w(s) = 0 \\[2mm]
w(s) = \nabla f(y(s)).
\end{cases}
\tag{67}
$$

The convergence rate becomes

$$
f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right) \text{ as } s \to +\infty.
\tag{68}
$$

So, we have accelerated the dynamic for the values, passing from the convergence rate $1/t$ to $1/s^2$. Moreover, by proceeding as in Subsection 3.1, we can return from (67) to the dynamics (65).

Let us now come with the averaging process. Given $s_0 > 0$, we attach to $y(\cdot)$ the new function $x : [s_0, +\infty[\to \mathcal{H}$ defined by

$$
\dot{x}(s) + \frac{\delta}{s}(x(s) - y(s)) = 0,
\tag{69}
$$

with $x(s_0) = x_0$ given in $\mathcal{H}$, and where δ is a positive coefficient to adjust. Equivalently

$$
y(s) = x(s) + \frac{s}{\delta}\dot{x}(s).
\tag{70}
$$

Let us formulate (67) in terms of x by eliminating y. We first obtain

$$
\begin{cases}
\lambda\left(\dot{x}(s) + \frac{s}{\delta}\ddot{x}(s) + \frac{1}{\delta}\dot{x}(s)\right) + \dot{w}(s) + \gamma s w(s) = 0 \\[2mm]
w(s) = \nabla f\left(x(s) + \frac{s}{\delta}\dot{x}(s)\right).
\end{cases}
\tag{71}
$$

After reduction we obtain

$$
\begin{cases}
\ddot{x}(s) + \frac{\delta+1}{s}\dot{x}(s) + \frac{\delta}{\lambda s}\dot{w}(s) + \frac{\delta\gamma}{\lambda}w(s) = 0 \\[2mm]
w(s) = \nabla f\left(x(s) + \frac{s}{\delta}\dot{x}(s)\right).
\end{cases}
\tag{72}
$$

Take $\delta = \alpha - 1$, $\gamma = \frac{\lambda}{\alpha-1}$ in (72). We have $\delta\gamma = \lambda$, which gives

$$
\begin{cases}
\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \frac{\alpha-1}{\lambda s}\dot{w}(s) + w(s) = 0 \\[2mm]
w(s) = \nabla f\left(x(s) + \frac{s}{\alpha-1}\dot{x}(s)\right).
\end{cases}
\tag{73}
$$

According to the general properties of the averaging process

$$y(s) = x(s) + \frac{s}{\alpha - 1}\dot{x}(s), \tag{74}$$

which has been already studied, see (30), we obtain the following theorem for a dynamics which combines explicit and implicit Hessian driven damping in a new way.

THEOREM 6.   *Suppose that $f \colon \mathcal{H} \to \mathbb{R}$ satisfies $(\mathcal{A})$. Let $x \colon [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system*

$$\begin{cases} \ddot{x}(s) + \dfrac{\alpha}{s}\dot{x}(s) + \dfrac{\alpha - 1}{\lambda s}\dot{w}(s) + w(s) = 0 \\[2mm] w(s) = \nabla f\left(x(s) + \dfrac{s}{\alpha - 1}\dot{x}(s)\right). \end{cases} \tag{75}$$

Assume that $\alpha > 1$. Then the following statements are true:

(i)   *(integral estimate of the gradients)* $\displaystyle\int_{s_0}^{+\infty} s^3 \|w(s)\|^2 \, ds < +\infty$;

(ii)   *(convergence of the gradients towards zero)* $\|w(s)\| = o\left(\frac{1}{s^2}\right)$ *as $s \to +\infty$*;

(iii)   *(convergence of the velocities towards zero)* $\|\dot{x}(s)\| = o\left(\frac{1}{s}\right)$ *as $s \to +\infty$*;

(iv)   *the solution trajectory $x(s)$ converges weakly as $s \to +\infty$, and its limit belongs to $S = \arg\min f$.*

If $\alpha > 3$, then

(v)   *(convergence of the values towards the minimal value)* $f(x(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right)$ *as $s \to +\infty$*.

Let $x(\cdot)$ be a solution trajectory of (75) which satisfies the Cauchy data $x(s_0) = x_0, \dot{x}(s_0) = x_1$, and $w(s_0) := \nabla f\left(x_0 + \frac{s_0}{\alpha - 1}x_1\right)$. In the same way as in the proof of Theorem 2, we first show that such a solution is reached by considering first the solution $y(\cdot)$ of the system

$$\begin{cases} \lambda \dot{y}(s) + \dot{w}(s) + \frac{\lambda s}{\alpha - 1}w(s) = 0 \\[2mm] w(s) = \nabla f(y(s)) \\[2mm] y(s_0) = x_0 + \frac{s_0}{\alpha - 1}x_1, \end{cases} \tag{76}$$

and then

$$\begin{cases} \dot{x}(s) + \frac{\alpha - 1}{s}(x(s) - y(s)) = 0 \\[2mm] x(t_0) = x_0. \end{cases} \tag{77}$$

Indeed, (77) gives $\dot{x}(s_0) = \frac{\alpha - 1}{s_0}(y(s_0) - x(s_0))$, which, by the second equation of (76), equals to x_1. We have already seen that x can be seen as an averaging process (77) of y as follows

$$x(s) = \int_{s_0}^{s} y(u) \, d\mu_s(u) + \xi(s), \tag{78}$$

24

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

where μ_s is the measure on $[s_0, s]$ defined by

$$\mu_s = \frac{s_0^{\alpha-1}}{s^{\alpha-1}}\delta_{s_0} + (\alpha-1)\frac{u^{\alpha-2}}{s^{\alpha-1}}du,$$

where δ_{s_0} is the Dirac measure at s_0 and

$$\xi(s) := -\frac{s_0^{\alpha}}{(\alpha-1)s^{\alpha-1}}x_1.$$

The statements (i) and (ii) follow from Theorem 5 after time rescaling and integration by substitution. In order to show (iv), we need to pass from the convergence of y to that of x by using the interpretation of x as an average of y plus a negligible term. Moreover, we know that y converges weakly to an element in S thanks to Theorem 5. The convergence of x follows from similar arguments as in Theorem 2.

Finally, since the averaging process is the same as in Theorem 2, the convergence rate of $1/s^2$ for the values (v) follows by the same arguments.   Q.E.D.

3.3. General Hessian damping coefficient: extension to bilevel convex optimization. Given β_0 a positive damping coefficient, consider the more general form of the dynamic

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \beta_0\frac{s}{\alpha-1}\dot{x}(s)\right) = 0. \tag{79}$$

Our study in the previous subsections concerned the case $\beta_0 = 1$. This section shows that the time rescaling and averaging technique links the dynamical system (79) with a bilevel convex optimization problem and thus emphasizes the versatility of the approach proposed in this paper.

Let us introduce

$$y(s) := x(s) + \beta_0\frac{s}{\alpha-1}\dot{x}(s), \tag{80}$$

so that (79) is written equivalently

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f(y(s)) = 0. \tag{81}$$

Let us reformulate (81) as a differential equation with $y(s)$ as the state variable. According to (80) we have

$$\dot{x}(s) = \frac{\alpha-1}{\beta_0 s}(y(s) - x(s)), \tag{82}$$

which after derivation gives

$$\ddot{x}(s) = \frac{\alpha-1}{\beta_0 s}(\dot{y}(s) - \dot{x}(s)) - \frac{\alpha-1}{\beta_0 s^2}(y(s) - x(s)). \tag{83}$$

Combining (81) and (83) we obtain

$$\frac{\alpha - 1}{\beta_0 s}(\dot{y}(s) - \dot{x}(s)) - \frac{\alpha - 1}{\beta_0 s^2}(y(s) - x(s)) + \frac{\alpha}{s}\dot{x}(s) + \nabla f(y(s)) = 0. \tag{84}$$

By replacing $\dot{x}(s)$ by its formulation given in (82) we obtain

$$\frac{\alpha - 1}{\beta_0 s}\dot{y}(s) + \left(\frac{\alpha}{s} - \frac{\alpha - 1}{\beta_0 s}\right)\frac{\alpha - 1}{\beta_0 s}(y(s) - x(s)) - \frac{\alpha - 1}{\beta_0 s^2}(y(s) - x(s)) + \nabla f(y(s)) = 0.$$

After reduction we get

$$\frac{\alpha - 1}{\beta_0 s}\dot{y}(s) + \frac{(\alpha - 1)^2(\beta_0 - 1)}{\beta_0^2 s^2}(y(s) - x(s)) + \nabla f(y(s)) = 0. \tag{85}$$

Equivalently

$$\dot{y}(s) + \frac{\beta_0 s}{\alpha - 1}\nabla f(y(s)) + \frac{(\alpha - 1)(\beta_0 - 1)}{\beta_0 s}(y(s) - x(s)) = 0. \tag{86}$$

Putting together (82) and (86) we obtain the system

$$\begin{cases} \dot{y}(s) + \dfrac{\beta_0 s}{\alpha - 1}\nabla f(y(s)) + \dfrac{(\alpha - 1)(\beta_0 - 1)}{\beta_0 s}(y(s) - x(s)) = 0 \\ \dot{x}(s) + \dfrac{\alpha - 1}{\beta_0 s}(x(s) - y(s)) = 0. \end{cases} \tag{87}$$

Let us now consider time scaling of the above system as in (21). Set

$$s = \sqrt{2(\alpha - 1)t} \text{ and } y(\sqrt{2(\alpha - 1)t}) = Y(t), \ x(\sqrt{2(\alpha - 1)t}) = X(t).$$

A similar calculation as in Subsection 2.2 gives

$$\begin{cases} \dot{Y}(t) + \beta_0 \nabla f(Y(t)) + \dfrac{(\alpha - 1)(\beta_0 - 1)}{2\beta_0 t}(Y(t) - X(t)) = 0 \\ \dot{X}(t) + \dfrac{\alpha - 1}{2\beta_0 t}(X(t) - Y(t)) = 0. \end{cases} \tag{88}$$

This is a perturbation of the steepest descent dynamical system in the product space $\mathcal{H} \times \mathcal{H}$. Precisely, setting $Z(t) = (Y(t), X(t)) \in \mathcal{H} \times \mathcal{H}$, we have

$$\dot{Z}(t) + \beta_0 \nabla \Psi(Z(t)) + \frac{(\alpha - 1)}{2\beta_0 t}A(Z(t)) = 0, \tag{89}$$

where $\Psi(Z) = f(Y)$ and $A : \mathcal{H} \times \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ is the linear operator whose matrix is given by

$$A = \begin{pmatrix} \beta_0 - 1 & -(\beta_0 - 1) \\ -1 & +1 \end{pmatrix}$$

26

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

The case $\beta_0 = 2$ is of particular interest since then A is a symmetric operator which is positive semidefinite. The study of this system then follows from classical results concerning gradient dynamical systems with multiscale aspects, see for example the work of Attouch and Czarnecki [15].

The dynamical system (89) for $\beta_0 = 2$ fits in the dynamics

$$\dot{Z}(t) + \beta_0 \nabla \Psi(Z(t)) + \varepsilon(t) \nabla \Phi(Z(t)) = 0, \tag{90}$$

with $\varepsilon(t) := \frac{(\alpha - 1)}{2\beta_0 t}$ and $\Phi(\cdot) = \frac{1}{2} \|\cdot\|_A^2$, where $\|\cdot\|_A$ is the seminorm induced by the symmetric and positive semidefinite operator A. The differential equation (90) can be seen as a continuous time approach of the bilevel optimization problem

$$\min_{\mathcal{H} \times \mathcal{H}} \left\{ \Phi(Z) \mid Z \in D := \operatorname*{arg\,min}_{\mathcal{H} \times \mathcal{H}} \Psi \right\}. \tag{91}$$

Let Z_* be a solution of (91), meaning that $\Phi(Z_*) = \inf_D \Phi$ and $\Psi(Z_*) = \inf_{\mathcal{H} \times \mathcal{H}} \Psi$. Define for $t \geq t_0$

$$\Psi_\Phi(t) := \beta_0 (\Psi(Z(t)) - \inf_{\mathcal{H} \times \mathcal{H}} \Psi) + \varepsilon(t) (\Phi(Z(t)) - \inf_D \Phi) \geq 0.$$

Indeed the above quantity is nonnegative because we are in the particular simple case where $\Phi(Y, X) = \frac{1}{2}\|Y - X\|^2$ is nonnegative, $D = \operatorname*{arg\,min}_{\mathcal{H}} f \times \mathcal{H}$ and hence $\inf_D \Phi = 0$. Let us compute

$$\begin{aligned}
\frac{d}{dt}\left(\frac{1}{2}\|Z(t) - Z_*\|^2\right) &= \langle Z(t) - Z_*, \dot{Z}(t) \rangle \\
&= -\beta_0 \langle Z(t) - Z_*, \nabla \Psi(Z(t)) \rangle - \varepsilon(t) \langle Z(t) - Z_*, \nabla \Phi(Z(t)) \rangle \\
&\leq -\beta_0 (\Psi(Z(t)) - \inf_{\mathcal{H} \times \mathcal{H}} \Psi) - \varepsilon(t) (\Phi(Z(t)) - \inf_D \Phi) \\
&= -\Psi_\Phi(t).
\end{aligned}$$

This shows that $\lim_{t \to +\infty} \|Z(t) - Z_*\| \in \mathbb{R}$ exists and that

$$\int_{t_0}^{+\infty} \Psi_\Phi(t)\, dt < +\infty.$$

This yields $\liminf_{t \to +\infty} t\Psi_\Phi(t) = 0$. On the other hand, we have

$$\begin{aligned}
\frac{d}{dt}\Psi_\Phi(t) &= \beta_0 \langle \nabla \Psi(Z(t)), \dot{Z}(t) \rangle + \varepsilon(t) \langle \nabla \Phi(Z(t)), \dot{Z}(t) \rangle + \dot{\varepsilon}(t)\left(\Phi(Z(t)) - \inf_D \Phi\right) \\
&\leq -\|\dot{Z}(t)\|^2.
\end{aligned}$$

This means that Ψ_Φ is decreasing and furthermore

$$\frac{d}{dt}\left(t\Psi_\Phi\left(t\right)\right) = \Psi_\Phi\left(t\right) + t\frac{d}{dt}\Psi_\Phi\left(t\right) \le \Psi_\Phi\left(t\right) - t\left\|\dot{Z}\left(t\right)\right\|^2.$$

Since $\Psi_\Phi \in \mathbb{L}^1\left(\left[t_0, +\infty\right[\right)$, we conclude that $t\|\dot{Z}\left(t\right)\|^2 \in \mathbb{L}^1\left(\left[t_0, +\infty\right[\right)$ and $\lim_{t\to+\infty} t\Psi_\Phi\left(t\right) = 0$. Consequently,

$$\lim_{t\to+\infty} t\left(\Psi\left(Z(t)\right) - \inf_{\mathcal{H}\times\mathcal{H}} \Psi\right) = 0 \text{ and } \lim_{t\to+\infty} \left(\Phi\left(Z(t)\right) - \inf_D \Phi\right) = 0.$$

Since Ψ and Φ are convex and lower semicontinuous, the second condition of Opial's lemma is fulfilled. Thus $Z(t)$ converges weakly to a solution of (91) as $t \to +\infty$.

Moreover, according to the definition of Φ and A, it holds

$$\Phi\left(Z(t)\right) = \frac{1}{2}\left\langle Z(t), A\left(Z\left(t\right)\right)\right\rangle = \frac{1}{2}\|X(t) - Y(t)\|^2 = \frac{2\beta_0^2}{(\alpha - 1)^2}t^2\left\|\dot{X}(t)\right\|^2 \to 0 \text{ as } t \to +\infty.$$

According to the definitions of Ψ and Φ we conclude that there exists $x_* \in \arg\min_{\mathcal{H}} f$ such that $(Y(t), X(t))$ converges weakly to (x_*, x_*) as $t \to +\infty$,

$$f(Y(t)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{t}\right) \text{ and } \lim_{t\to+\infty}\|X(t) - Y(t)\| = \lim_{t\to+\infty} t\|\dot{X}(t)\| = 0.$$

This means that $x(s)$ converges weakly to x_* as $s \to +\infty$,

$$f\left(x(s) + \frac{2s}{\alpha - 1}\dot{x}(s)\right) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right) \text{ and } \lim_{s\to+\infty} s\|\dot{x}(s)\| = 0.$$

In the following theorem we collect the convergence properties of the trajectory of the dynamical system (79) in case $\beta_0 = 2$.

THEOREM 7.   *Suppose that $f: \mathcal{H} \to \mathbb{R}$ satisfies $(\mathcal{A})$. Let $x: [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system*

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f\left(x(s) + \frac{2s}{\alpha - 1}\dot{x}(s)\right) = 0.$$

Assume that $\alpha > 1$. Then the following statements are true:

(i) *(convergence of the values towards the minimal value)* $f\left(x(s) + \frac{2s}{\alpha-1}\dot{x}(s)\right) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right)$ *as $s \to +\infty$;*

(ii) *(convergence of the velocities towards zero)* $\|\dot{x}(s)\| = o\left(\frac{1}{s}\right)$ *as $s \to +\infty$;*

(iii) *the solution trajectory $x(s)$ converges weakly as $s \to +\infty$, and its limit belongs to $S = \arg\min f$.*

28

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

3.4. From second-order to third-order accelerated systems.   Let us develop the time-scale and averaging technique, starting from a second-order evolution system. According to the general properties of this process, this will provide a third-order in time evolution system. This section is inspired by recent articles by Attouch-Chbani-Riahi [13], [14], who first introduced fast third-order damped inertial systems by considering the time-scale technique. In fact, we will show that these results fall within the general framework that we have developed in the previous sections. As a starting point for our study, consider the second-order dynamic with Asymptotic Vanishing Damping

$$(\text{AVD})_\alpha \quad \ddot{z}(t) + \frac{\alpha}{t}\dot{z}(t) + \nabla f(z(t)) = 0$$

which was introduced by Su-Boyd-Candès [48]. The importance of this dynamic comes from the fact that the accelerated gradient method of Nesterov can be obtained as a temporal discretization by taking $\alpha = 3$. Let us briefly recall the convergence properties of this system:

• For $\alpha \geq 3$, each trajectory $z(\cdot)$ of $(\text{AVD})_\alpha$ satisfies the asymptotic convergence rate of the values $f(z(t)) - \inf_{\mathcal{H}} f = O\left(1/t^2\right)$ as $t \rightarrow +\infty$, see [6], [11], [39], [48].

• For $\alpha > 3$, it has been shown in [11] that each trajectory converges weakly to a minimizer. The corresponding algorithmic result has been obtained by Chambolle-Dossal [35]. In addition, it is shown in [20] and [39] that the asymptotic convergence rate of the values is $o(1/t^2)$.

These rates are optimal, that is, they can be reached, or approached arbitrarily close. For further results concerning the system $(\text{AVD})_\alpha$ one can consult [6, 7, 11, 20, 39, 48].

Let's make the time rescaling of $(\text{AVD})_\alpha$ given by

$$t = \tau(s) = \frac{2}{3}s^{\frac{3}{2}}, \quad y(s) = z(\tau(s)).$$

According to the classical derivation chain rule we first obtain

$$\ddot{y}(s) + \frac{\alpha\dot{\tau}(s)^2 - \tau(s)\ddot{\tau}(s)}{\tau(s)\dot{\tau}(s)}\dot{y}(s) + \dot{\tau}(s)^2\nabla f(y(s)) = 0. \tag{92}$$

Then replacing $\tau(s)$ by its expression $\tau(s) = \frac{2}{3}s^{\frac{3}{2}}$, we obtain the rescaled dynamic

$$\ddot{y}(s) + \frac{3\alpha - 1}{2s}\dot{y}(s) + s\nabla f(y(s)) = 0. \tag{93}$$

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

29

Since $\alpha > 3$ is equivalent to $\frac{3\alpha - 1}{2} > 4$, we obtain from here that, for $\alpha > 3$, for any solution trajectory $y(\cdot)$ of

$$\ddot{y}(s) + \frac{\alpha + 1}{s} \dot{y}(s) + s \nabla f(y(s)) = 0, \tag{94}$$

we have

$$f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^3}\right) \quad \text{as } s \to +\infty, \tag{95}$$

and that the trajectory $y(s)$ converges weakly to an element in $\inf_{\mathcal{H}} f$ as s tends to $+\infty$. Let us now proceed with the averaging process, and introduce the new variable $x(\cdot)$ defined by

$$\frac{1}{4} s \dot{x}(s) + x(s) = y(s).$$

Then

$$\dot{y}(s) = \frac{1}{4} s \ddot{x}(s) + \frac{5}{4} \dot{x}(s) \quad \text{and} \quad \ddot{y}(s) = \frac{1}{4} s \dddot{x}(s) + \frac{3}{2} \ddot{x}(s).$$

As a consequence, (94) becomes the following third-order evolution system where we have replaced $4f$ by f, which does not affect the convergence rates:

$$\dddot{x}(s) + \frac{\alpha + 7}{s} \ddot{x}(s) + \frac{5(\alpha + 1)}{s^2} \dot{x}(s) + \nabla f\left(x(s) + \frac{1}{4} s \dot{x}(s)\right) = 0.$$

According to the general properties of the averaging process, which preserves the convergence rate of the values, we obtain the following result.

THEOREM 8.   *Let $x : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the evolution system*

$$\dddot{x}(s) + \frac{\alpha + 7}{s} \ddot{x}(s) + \frac{5(\alpha + 1)}{s^2} \dot{x}(s) + \nabla f\left(x(s) + \frac{1}{4} s \dot{x}(s)\right) = 0. \tag{96}$$

Suppose that $\alpha > 3$. Then, as $s \to +\infty$

(i) $f(x(s)) - \inf_{\mathcal{H}} f = o\left(\dfrac{1}{s^3}\right);$

(ii) *the trajectory converges weakly as $s \to +\infty$, that is $x(s) \rightharpoonup x_\infty$, and $x_\infty \in \arg\min_{\mathcal{H}} f$.*

Proof.   Let $x : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the evolution system (96) which satisfies the Cauchy data $x(s_0) = x_0, \dot{x}(s_0) = x_1$, and $\ddot{x}(s_0) = x_2$. Let us verify that this solution comes from the dynamic

$$\begin{cases} \ddot{y}(s) + \dfrac{\alpha + 1}{s} \dot{y}(s) + s \nabla f(y(s)) = 0 \\ y(s_0) = x_0 + \dfrac{s_0}{4} x_1 \\ \dot{y}(s_0) = \dfrac{1}{4} s_0 x_2 + \dfrac{5}{4} x_1 \end{cases} \tag{97}$$

then by averaging

$$\begin{cases} \dot{x}(s) + \frac{4}{s}(x(s) - y(s)) = 0 \\ \\ x(s_0) = x_0. \end{cases} \tag{98}$$

The first equation in (98) gives $\dot{x}(s_0) = \frac{4}{s_0}(y(s_0) - x(s_0))$, which, by the second equation of (97), gives $\dot{x}(s_0) = x_1$. On the other hand, taking time derivative of yields $\dot{y}(s_0) = \frac{1}{4}s_0\ddot{x}(s_0) + \frac{5}{4}\dot{x}(s_0)$. Taking into account the third equation of (97), and that $\dot{x}(s_0) = x_1$ as we just showed, we get $\ddot{x}(s_0) = x_2$.

Multiplying both sides of the first equation in (98) by s^4, we get

$$\frac{d}{ds}\left(s^4 x(s)\right) = 4s^3 y(s). \tag{99}$$

Integration of (99) from s_0 to s, then using the data $x(s_0) = x_0$ and $y(s_0) = x_0 + \frac{s_0}{4}x_1$, we obtain,

$$x(s) = \frac{s_0^4}{s^4}x_0 + \frac{4}{s^4}\int_{s_0}^{s} r^3 y(r)\, dr = \frac{s_0^4}{s^4}y(s_0) + \frac{4}{s^4}\int_{s_0}^{s} r^3 y(r)\, dr - \frac{s_0^5}{4s^4}x_1.$$

Let μ_s be the measure on $[s_0, s]$ defined by

$$\mu_s = \frac{s_0^4}{s^4}\delta_{t_0} + \frac{4r^3}{s^4}dr,$$

where δ_{t_0} is the Dirac measure at t_0, and set

$$\xi(s) := -\frac{s_0^5}{4s^4}x_1$$

Then, observe that $x(s)$ can be written as follows

$$x(s) = \int_{s_0}^{s} y(r)d\mu_s(r) + \xi(s). \tag{100}$$

Since $y(s)$ converges weakly as $s \to +\infty$, we have that the trajectory $y(\cdot)$ remains bounded. From (34) we deduce that $x(\cdot)$ is also bounded. Since ∇f is Lipschitz continuous on the bounded sets, it follows that ∇f is bounded on the bounded sets, which immediately implies that f is also Lipschitz continuous on the bounded sets. Therefore

$$f(x(s)) - f\left(\int_{s_0}^{s} y(r)d\mu_s(r)\right) \le C\|w(s)\| \le \frac{C}{s^4}. \tag{101}$$

On the other hand, we have from (95) that there exists a positive function $\varepsilon(\cdot)$ such that $\lim_{s\to+\infty} \varepsilon(s) = 0$ and

$$f(y(s)) - \inf_{\mathcal{H}} f = \frac{\varepsilon(s)}{s^3}$$

This, together with the convexity of f and **Jensen's inequality**, allow us to deduce that for every $s \geq s_0$

$$f\left(\int_{s_0}^s y(r)d\mu_s(r)\right) - \inf_{\mathcal{H}} f \leq \int_{s_0}^s (f(y(r)) - \inf_{\mathcal{H}} f)\, d\mu_s(r) = \int_{s_0}^s \frac{\varepsilon(s)}{r^3} d\mu_s(r).$$

According to the definition of μ_s, we deduce that

$$f(x(s)) - \inf_{\mathcal{H}} f = \frac{s_0^4}{s^4} + \frac{4}{s^4} \int_{s_0}^s \varepsilon(r)dr. \tag{102}$$

Combining (101) with (102) we obtain the fast convergence of the values, as

$$\limsup_{s\to+\infty} s^3 (f(x(s)) - \inf_{\mathcal{H}} f) \leq \limsup_{s\to+\infty} \left(\frac{C}{s} + \frac{4}{s} \int_{s_0}^s \varepsilon(r)dr\right), \tag{103}$$

and the first statement follows from Lemma 4.

The proof of the second statement is similar to the one given in Theorem 2. Q.E.D.

REMARK 6. Given that the coefficient in front of the gradient in the dynamic (96) is fixed, one can anticipate deriving associated gradient algorithms with similar convergence rates through explicit discretization, namely $o(1/k^3)$ as $k \to +\infty$. Further research will be dedicated to elucidating the discrepancy between the continuous-time convergence rate of $o(1/t^3)$ and the lower complexity bound of $1/k^2$ for first-order algorithms in the minimization of convex differentiable functions, as demonstrated by Nesterov (refer to [41, Theorem 2.1.7]). Explicit discretization of (96) would necessitate the evaluation of the function gradient at $x_k + \frac{k}{4}(x_k - x_{k-1})$. It's worth noting that this extrapolation is reasonable, given that, according to classical estimates of the velocity vector, we have $\lim_{s\to+\infty} s\dot{x}(s) = 0$.

4. The nonsmooth case. In this section, we show that the approach proposed in this paper extends to the case when $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is assumed to be a (possibly nonsmooth) proper, convex and lower semicontinuous function.

4.1. The continuous dynamics. As a basic property of the dynamical system (28), the couple of variables (x, y) defined in Subsection 2.3 satisfies the following differential system, which involves only first-order derivatives in time and space

$$\begin{cases} \dot{y}(s) + \frac{s}{\alpha-1}\nabla f(y(s)) & = 0 \\ \dot{x}(s) + \frac{\alpha-1}{s}(x(s) - y(s)) & = 0. \end{cases} \tag{104}$$

This naturally suggests the extension of the above results to the nonsmooth case (replace the gradient of f by its subdifferential). We thus get the differential inclusion system

$$\begin{cases} \dot{y}(s) + \frac{s}{\alpha-1}\partial f(y(s)) & \ni 0 \\ \dot{x}(s) + \frac{\alpha-1}{s}(x(s) - y(s)) & = 0. \end{cases} \tag{105}$$

Solving this system gives generalized solutions to the second-order differential inclusion

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \partial f\left(x(s) + \frac{s}{\alpha-1}\dot{x}(s)\right) \ni 0 \tag{106}$$

whose direct study raises several difficulties, recalling that $x(s_0) = x_0$ and $\dot{x}(s_0) = x_1$. To extend the results of the previous section to this nonsmooth case, we need to avoid the arguments using the Lipschitz continuity of ∇f. So, we are led to consider the Cauchy problem for (106) with initial data $x_0 \in \text{dom} f$ and $x_1 = 0$, that is with initial velocity equal to zero. So doing we have $y(s_0) = x(s_0) = x_0$, which allows to interpret x as an average of y without correcting term ($\xi = 0$ with the notations of Subsection 2.6).

The existence and uniqueness of a strong solution to the associated Cauchy problem relies on the equivalent formulation of (105) as a perturbation of the generalized steepest descent dynamical system in the product space $\mathcal{H} \times \mathcal{H}$. Precisely, define $F : \mathcal{H} \times \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, for any $Z = (y, x) \in \mathcal{H} \times \mathcal{H}$ by

$$F(Z) = f(y),$$

and let $G : \mathcal{H} \times \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ be the operator defined by

$$G(Z) = (0, x - y). \tag{107}$$

Then (105) is written equivalently as

$$\dot{Z}(s) + \frac{s}{\alpha-1}\partial F(Z(s)) + \frac{\alpha-1}{s}G(Z(s)) \ni 0. \tag{108}$$

The initial condition becomes $Z(t_0) = (x_0, x_0)$ which belongs to dom $F =$ dom $f \times \mathcal{H}$. According to the classical results concerning the Lipschitz perturbation of evolution equations governed by subdifferentials of convex functions, see [31, Proposition 3.12], we obtain the existence and uniqueness of a global strong solution of the Cauchy problem associated with (105). As a major advantage of the time scaling and averaging techniques, the arguments used in the previous section still work in this more general nonsmooth situation. The rules of differential calculus are still valid for strong solutions, see [32, chapter VIII.2], and Jensen's inequality is still valid for a nonsmooth function f (see e.g. [27, Proposition 9.24]). Indeed Jensen's inequality is classical for a smooth convex function f. Its extension to the nonsmooth convex case can be obtained by first writing it for the Moreau-Yosida regularization f_λ of f, then passing to the limit when $\lambda \downarrow 0$. According to the monotone convergence of f_λ towards f, we can pass to the limit in the integral term thanks to the Beppo Levi monotone convergence theorem. Therefore we obtain the following theorem.

THEOREM 9.   *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous, and convex function such that $S = \arg\min f \neq \emptyset$. Let $x : [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of*

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \partial f\left(x(s) + \frac{s}{\alpha - 1}\dot{x}(s)\right) \ni 0, \tag{109}$$

which satisfies the initial conditions $x(s_0) \in$ dom f and $\dot{x}(s_0) = 0$. Assume that $\alpha > 1$. Then

(i) *the solution trajectory $x(s)$ converges weakly as $s \to +\infty$, and its limit belongs to $S = \arg\min f$.*

   If $\alpha > 3$, then

(ii) *(convergence of the values towards the minimal value) $f(y(s)) - \inf_{\mathcal{H}} f = o\left(\frac{1}{s^2}\right)$ as $s \to +\infty$.*

REMARK 7.   As a consequence of (ii), we have that $x(s)$ remains in the domain of f for all $s \geq t_0$. This viability property strongly depends on the fact that the initial position belongs to the domain of f, and that the initial velocity has been taken equal to zero.

REMARK 8.   The above theorem is valid in an infinite dimensional setting, which makes it applicable to nonlinear PDEs. According to the classical theory for the steepest descent dynamic in the nonsmooth convex case, we have that at each $s > s_0$ the right derivative of $y(s)$ exists and satisfies

$$-\left(\frac{dy}{ds}\right)^+(s) = \frac{s}{\alpha - 1}\left(\partial f(y(s))\right)^0$$

where $(\partial f(y(s)))^0$ is the element of minimum norm of $\partial f(y(s))$. Thus, y may exhibit shocks. By contrast, x, which is an average of y, has a continuous derivative, hence does not exhibit shocks. Still x is not twice differentiable.

34

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

4.2. A proximal type algorithm for nonsmooth minimization.

Recall the second-order differential inclusion (106) associated with (1) in Section 4

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \partial f\left(x(s) + \frac{s}{\alpha - 1}\dot{x}(s)\right) \ni 0, \tag{110}$$

or equivalently formulated as a first order system (see (105))

$$\begin{cases} \dot{y}(s) + \frac{s}{\alpha-1}\partial f(y(s)) & \ni 0 \\[2mm] \dot{x}(s) + \frac{\alpha-1}{s}(x(s) - y(s)) & = 0. \end{cases} \tag{111}$$

We consider the following implicit discretization in time of (111), written for every $k \geq 0$ as

$$\begin{cases} y_{k+1} - y_k + \frac{s_{k+1}}{\alpha-1}\partial f(y_{k+1}) & \ni 0 \\[2mm] x_{k+1} - x_k + \frac{\alpha-1}{s_{k+1}}(x_k - y_{k+1}) & = 0. \end{cases} \tag{112}$$

This gives rise to the following numerical algorithm

$$(\forall k \geq 0) \quad \begin{cases} y_{k+1} & := \operatorname{prox}_{\frac{s_{k+1}}{\alpha-1}f}(y_k) \\[2mm] x_{k+1} & := \left(1 - \frac{\alpha-1}{s_{k+1}}\right)x_k + \frac{\alpha-1}{s_{k+1}}y_{k+1}, \end{cases} \tag{113}$$

where $y_0, x_0 \in \mathcal{H}$ are arbitrary initial points. This algorithm has some analogy with the relaxed inertial proximal algorithm for maximally monotone operators in [8].

For the step size sequence $(s_k)_{k\geq 0}$ we impose the following recurrence condition

$$s_0 := 0 \qquad \text{and} \qquad s_{k+1}^2 - (\alpha - 1)s_{k+1} = s_k^2 \quad \forall k \geq 0, \tag{114}$$

which relates to the Nesterov step size rule. Indeed, by denoting $t_k := \frac{s_k}{\alpha-1}$, we have for every $k \geq 0$

$$s_{k+1} := \frac{\alpha - 1 + \sqrt{(\alpha - 1)^2 + 4s_k^2}}{2} \quad \text{or, equivalently,} \quad t_{k+1} := \frac{1 + \sqrt{1 + 4t_k^2}}{2},$$

which is nothing but the classic step size rule of Nesterov's accelerated gradient method [26, 40, 41].

We will nevertheless continue to work with the sequence $(s_k)_{k\geq 0}$ in order to emphasize that all the convergence statements provided in this section are valid for $\alpha > 1$, contrary to the stronger condition $\alpha > 3$ for the low resolution case (see Remark 1). We have that $s_k \sim k$, precisely, $s_k \geq (k+1)(\alpha - 1)/2$ for every $k \geq 0$ (see [26, Lemma 4.3]). Furthermore, by a telescoping sum argument, we have from (114)

$$s_{k+1}^2 - (\alpha - 1)\sum_{i=0}^{k} s_{i+1} = s_0^2.$$

Since $s_0 = 0$, we can conclude that for every $k \geq 0$

$$s_k^2 = (\alpha - 1) \sum_{i=0}^{k} s_i. \tag{115}$$

Let us also notice that (113) can be written only in terms of the sequence $(x_k)_{k \geq 0}$

$$x_{k+1} - x_k - \frac{s_k - (\alpha - 1)}{s_{k+1}} (x_k - x_{k-1}) + \partial f \left(x_k + \frac{s_{k+1}}{\alpha - 1} (x_{k+1} - x_k) \right) \ni 0 \quad \forall k \geq 1,$$

which can be seen as a direct discretization of (110).

In the following we first analyze the convergence properties of the sequence $(y_k)_{k \geq 0}$ generated by (113). Then, we transfer these to the sequence $(x_k)_{k \geq 0}$. One can easily observe the similarity between this approach and the continuous time one, where we transferred the converge properties of the trajectory $y(\cdot)$ generated by the continuous steepest descent system after time scaling to the averaged trajectory $x(\cdot)$.

THEOREM 10. *Let $(y_k)_{k \geq 0}$ be the sequence generated by (113). The following statements are true:*

(i) *(summability of the function values)* $\quad \sum_{k \geq 0} s_k (f(y_k) - \inf_{\mathcal{H}} f) < +\infty$;

(ii) *(summability of the subgradients) there exists a sequence $(\eta_k)_{k \geq 0}$ such that $\eta_k \in \partial f(y_k)$ for every $k \geq 0$ and $\sum_{k \geq 0} s_k^2 \|\eta_k\|^2 < +\infty$.*

In particular, if f is differentiable, then $\quad \|\nabla f(y_k)\| = o\left(\frac{1}{\sqrt{\sum_{i=0}^{k} s_i^2}} \right) = o\left(\frac{1}{k\sqrt{k}} \right)$ *as $k \to +\infty$*;

(iii) *(convergence of the values towards the minimal value)* $\quad f(y_k) - \inf_{\mathcal{H}} f = O\left(\frac{1}{\sum_{i=0}^{k} s_i} \right) = o\left(\frac{1}{s_k^2} \right) = o\left(\frac{1}{(k+1)^2} \right)$ *as $k \to +\infty$*;

(iv) *the sequence of iterates $(y_k)_{k \geq 0}$ converges weakly as $k \to +\infty$, and its limit belongs to $S = \arg\min_{\mathcal{H}} f$.*

Proof. Let $k \geq 0$ be fixed. Take $z_* \in S = \arg\min f$. According to (113) there exists $\eta_{k+1} \in \partial f(y_{k+1})$ such that $y_{k+1} - y_k + \frac{s_{k+1}}{\alpha - 1} \eta_{k+1} = 0$. By convexity of f we deduce that

$$
\begin{aligned}
\frac{1}{2} \|y_{k+1} - z_*\|^2 &= \frac{1}{2} \|y_k - z_*\|^2 + \langle y_{k+1} - z_*, y_{k+1} - y_k \rangle - \frac{1}{2} \|y_{k+1} - y_k\|^2 \\
&= \frac{1}{2} \|y_k - z_*\|^2 - \frac{s_{k+1}}{\alpha - 1} \langle y_{k+1} - z_*, \eta_{k+1} \rangle - \frac{1}{2} \frac{s_{k+1}^2}{(\alpha - 1)^2} \|\eta_{k+1}\|^2 \\
&\leq \frac{1}{2} \|y_k - z_*\|^2 - \frac{s_{k+1}}{\alpha - 1} (f(y_{k+1}) - \inf_{\mathcal{H}} f) - \frac{1}{2} \frac{s_{k+1}^2}{(\alpha - 1)^2} \|\eta_{k+1}\|^2. \tag{116}
\end{aligned}
$$

36

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

The summability statements in (i) and (ii) follow from [25, Lemma 5.31]. In addition, the limit $\lim_{k \to +\infty} \|y_k - z_*\| \in \mathbb{R}$ exists, which means that the first condition of the discrete Opial's lemma is fulfilled.

The convergence rate in (ii) follows from the fact that sequence $(\|\eta_k\|)_{k \geq 0}$ is nonincreasing. Indeed, we have from the monotonicity of ∂f

$$
\begin{aligned}
\|\eta_k\|^2 - \|\eta_{k+1}\|^2 &= -2 \langle \eta_{k+1}, \eta_{k+1} - \eta_k \rangle + \|\eta_{k+1} - \eta_k\|^2 \\
&= \frac{2(\alpha - 1)}{s_{k+1}} \langle y_{k+1} - y_k, \eta_{k+1} - \eta_k \rangle + \|\eta_{k+1} - \eta_k\|^2 \geq 0,
\end{aligned}
$$

and the rate follows from [7, Lemma 22]. Similarly, we notice that the sequence $(f(y_k) - \inf_{\mathcal{H}} f)_{k \geq 0}$ is nonincreasing as well. Precisely, for every $k \geq 0$ we have

$$
(f(y_k) - \inf_{\mathcal{H}} f) - (f(y_{k+1}) - \inf_{\mathcal{H}} f) \geq \langle \eta_{k+1}, y_k - y_{k+1} \rangle = \frac{\alpha - 1}{s_{k+1}} \|y_{k+1} - y_k\|^2 \geq 0.
$$

According to [7, Lemma 22] we get

$$
f(y_k) - \inf_{\mathcal{H}} f = o \left(\frac{1}{\sum_{i=0}^{k} s_i} \right),
$$

which proves (iii), thanks to (115). Finally, since $\lim_{k \to +\infty} f(y_k) = \inf_{\mathcal{H}} f$ and f is convex and lower semicontinuous, the second condition of Opial's lemma is also fulfilled. This gives the weak convergence of the sequence $(y_k)_{k \geq 0}$ to an element in $S = \arg\min f$.   Q.E.D.

Let us consider the convergence properties of $(x_k)_{k \geq 0}$ which are associated with the averaging process.

THEOREM 11.   *Let $(x_k)_{k \geq 0}$ be the sequence generated by* (113). *The following statements are true:*

(i) *(convergence of the values)* $f(x_k) - \inf_{\mathcal{H}} f = O \left(\frac{1}{\sum_{i=0}^{k} s_i} \right) = O \left(\frac{1}{s_k^2} \right) = O \left(\frac{1}{(k+1)^2} \right)$ *as* $k \to +\infty$;

(ii) *the sequence of iterates $(x_k)_{k \geq 0}$ converges weakly as $k \to +\infty$, and its limit belongs to $S = \arg\min_{\mathcal{H}} f$.*

Proof.   We begin by interpreting $(x_k)_{k \geq 0}$ as an average of $(y_k)_{k \geq 0}$. Multiplying the second equation in (113) by s_{k+1}^2 and using (114), we obtain for every $k \geq 0$

$$
s_{k+1}^2 x_{k+1} = \left(s_{k+1}^2 - (\alpha - 1) s_{k+1} \right) x_k + (\alpha - 1) s_{k+1} y_{k+1} = s_k^2 x_k + (\alpha - 1) s_{k+1} y_{k+1}.
$$

By a telescopic sum argument and using (115), we obtain from here the following expression for each $k \geq 0$

$$x_{k+1} = \frac{1}{s_{k+1}^2} \sum_{i=0}^{k+1} (\alpha - 1) s_i y_i = \frac{1}{\sum_{i=0}^{k+1} s_i} \sum_{i=0}^{k+1} s_i y_i.$$

Then, Jensen's inequality gives for every $k \geq 0$

$$f(x_{k+1}) - \inf_{\mathcal{H}} f = (f - \inf_{\mathcal{H}} f) \left(\frac{1}{\sum_{i=0}^{k+1} s_i} \sum_{i=0}^{k+1} s_i y_i \right)$$

$$\leq \frac{1}{\sum_{i=0}^{k+1} s_i} \sum_{i=0}^{k+1} s_i \left(f(y_i) - \inf_{\mathcal{H}} f \right) \leq \frac{\alpha - 1}{s_{k+1}^2} \sum_{i \geq 0} s_i \left(f(y_i) - \inf_{\mathcal{H}} f \right), \qquad (117)$$

where in the last equation we use (115). In Theorem 10 we have shown $\sum_{i \geq 0} s_i \left(f(y_i) - \inf_{\mathcal{H}} f \right) < +\infty$, which gives the announced convergence rates. The fact that $(x_k)_{k \geq 0}$ converges weakly to an element in $S = \arg \min f$ as $k \to +\infty$ follows from the fact that convergence entails ergodic convergence. Q.E.D.

5. The operator case.

What underlies our study is the first-order in time evolution system taken as a starting point and its convergence properties. In this regard, it is natural to consider also the evolution equation governed by a cocoercive operator. Indeed, gradients of convex functions and cocoercive operators are the two basic examples of monotone operators ensuring the convergence of the associated semi-groups of contractions. We will successively examine a general cocoercive operator, then the additively structured case. For monotone operators, the convergence rates are expressed in terms of velocities and residuals.

5.1. The general cocoercive case.

Recall that a single-valued operator $M : \mathcal{H} \to \mathcal{H}$ is called ρ-cocoercive (with $\rho > 0$) if

$$\langle M(x) - M(y), x - y \rangle \geq \rho \|M(x) - M(y)\|^2 \quad \forall x, y \in \mathcal{H}.$$

We mention the following classical properties of cocoercive operators:

a) If M is ρ-cocoercive, then it is monotone and $\frac{1}{\rho}$-Lipschitz continuous, and hence maximally monotone.

b) If f is a convex differentiable function whose gradient is L-Lipschitz continuous, then the operator ∇f is $1/L$-cocoercive (Baillon-Haddad theorem).

38

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

Let us come to our study. Given $M \colon \mathcal{H} \to \mathcal{H}$ a ρ-cocoercive operator, consider the monotone equation

$$M(x) = 0, \tag{118}$$

and the associated evolution equation

$$\dot{z}(t) + M(z(t)) = 0. \tag{119}$$

We denote the solution set of (118) by S, which represents the set of zeros $\operatorname{Zer} M$ of the operator M and is assumed to be nonempty. We recall the results in [27, Theorems 11 and 16] which state that $z(t)$ converges weakly to a point in S. In addition

$$\int_{t_0}^{+\infty} \|\dot{z}(t)\|^2 \, dt = \int_{t_0}^{+\infty} \|M(z(t))\|^2 \, dt < +\infty \text{ and } \|M(z(t))\| = o\left(\frac{1}{\sqrt{t}}\right) \text{ as } t \to +\infty. \tag{120}$$

Further we develop an analysis similar to the one in Subsection 3.1 in the case of a gradient operator. Set $y(s) := z(\tau(s))$, where $\tau \colon [t_0, +\infty[\to \mathbb{R}^+$ is a continuously differentiable increasing function satisfying $\lim_{s \to +\infty} \tau(s) = +\infty$. According to the derivation chain rule, we have

$$\dot{y}(s) = \dot{\tau}(s)\,\dot{z}(\tau(s))$$

which, by setting $t := \tau(s)$ in (119), gives

$$\dot{z}(\tau(s)) + M(z(\tau(s))) = 0.$$

Combining the two relations above gives

$$\dot{y}(s) + \dot{\tau}(s)\,M(y(s)) = 0. \tag{121}$$

Taking the derivative with respect to s of the above equation gives

$$\ddot{y}(s) + \ddot{\tau}(s)\,M(y(s)) + \dot{\tau}(s)\,\frac{d}{ds}(M(y(s))) = 0. \tag{122}$$

Multiplying both sides of (121) by $\frac{\alpha}{s}$, then adding the result to (122), we obtain

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \dot{\tau}(s)\,\frac{d}{ds}(M(y(s))) + \left(\ddot{\tau}(s) + \frac{\alpha}{s}\dot{\tau}(s)\right) M(y(s)) = 0. \tag{123}$$

As a subsequent result of (120) we have

$$\|M(y(s))\| = o\left(\frac{1}{\sqrt{\tau(s)}}\right) \text{ as } s \to +\infty.$$

Taking

$$\tau(s) = \frac{s^2}{2(\alpha+1)},$$

then (123) becomes

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \frac{s}{\alpha+1}\frac{d}{ds}(M(y(s))) + M(y(s)) = 0. \tag{124}$$

We obtain the rate

$$\|M(y(s))\| = o\left(\frac{1}{s}\right) \text{ as } s \to +\infty.$$

Moreover, it holds

$$\int_{\tau(t_0)}^{+\infty} s\,\|M(y(s))\|^2\,ds < +\infty.$$

Now let us do the reverse, and start from a solution trajectory $y(\cdot)$ of (123) satisfying $y(s_0) := y_0$ and $\dot{y}(s_0) := y_1$. Using similar arguments as in Theorem 4, we multiply both sides of (124) by $s^\alpha > 0$ and get

$$\frac{d}{ds}\left(s^\alpha \dot{y}(s) + \frac{s^{\alpha+1}}{\alpha+1}M(y(s))\right) = 0.$$

This leads to

$$\dot{y}(s) + \frac{s}{\alpha+1}M(y(s)) = \frac{c_0}{s^\alpha}, \tag{125}$$

where $c_0 \in \mathcal{H}$ is a constant that can be determined from the Cauchy condition. Set $z(t) := y(\sigma(t))$, where

$$\sigma(t) = \sqrt{2(\alpha+1)\,t}. \tag{126}$$

After calculations similar to those made previously, we arrive at

$$\dot{z}(t) + M(z(t)) = \frac{c}{t^{\frac{\alpha+1}{2}}},$$

where $c \in \mathcal{H}$ is a constant. Our study therefore falls within the properties of the perturbed system

$$\dot{z}(t) + M(z(t)) = g(t), \tag{127}$$

where, as $\alpha > 1$, the external perturbation $g : [t_0, +\infty[\to \mathcal{H}, g(t) = \frac{c}{t^{\frac{\alpha+1}{2}}}$, is such that

$$\int_{t_0}^{+\infty} \|g(t)\|\,dt < +\infty \text{ and } \int_{t_0}^{+\infty} t\,\|g(t)\|^2\,dt < +\infty. \tag{128}$$

We have the following result.

THEOREM 12. *Let $z : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of*

$$\dot{z}(t) + M(z(t)) = g(t), \tag{129}$$

where g fulfils (128). Then the following statements are true:

40

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

(i) *(convergence rate of the operator norm)* $\quad \|M(z(t))\| = o\left(\frac{1}{\sqrt{t}}\right)$ *as* $t \to +\infty$;

(ii) *(integral estimate of the operator norm)* $\quad \int_{t_0}^{+\infty} \|M(z(t))\|^2 \, dt < +\infty$;

(iii) *the solution trajectory $z(t)$ converges weakly as $t \to +\infty$, and its limit belongs to $S = \operatorname{Zer} M$.*

Combining the above theorem with the time scaling defined in (126) gives the following result.

THEOREM 13.　*Let $M : \mathcal{H} \to \mathcal{H}$ be a cocoercive operator and $y : [s_0, +\infty[\to \mathcal{H}$ be a solution trajectory of*

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \frac{s}{\alpha+1}\frac{d}{ds}(M(y(s))) + M(y(s)) = 0, \tag{130}$$

where $\alpha > 1$. Then the following statements are true:

(i) *(convergence rate of the operator norm)* $\quad \|M(y(s))\| = o\left(\frac{1}{s}\right)$ *as* $s \to +\infty$;

(ii) *(integral estimate of the gradients)* $\quad \int_{s_0}^{+\infty} s \|M(y(s))\|^2 \, ds < +\infty$;

(iii) *the solution trajectory $y(s)$ converges weakly as $s \to +\infty$, and its limit belongs to $S = \operatorname{Zer} M$.*

REMARK 9.　Theorem 13 is valid for an arbitrary cocoercive operator. By contrast, the corresponding result without the Newton correction term requires that the coefficient of the cocoercive operator asymptotically tends to infinity, see Attouch and Peypouquet [21].

REMARK 10.　Recently, in [29] the closely related dynamics

$$\ddot{y}(s) + \frac{\alpha}{s}\dot{y}(s) + \frac{2s}{\alpha+1}\frac{d}{ds}(M(y(s))) + M(y(s)) = 0 \tag{131}$$

has been considered. Compared to (124), the difference lies in the coefficient of $\frac{d}{ds}(M(y(s)))$. According to [29, Theorem 4 and Theorem 7], the solution trajectory of (131) verifies

$$\int_{t_0}^{+\infty} s^3 \|M(y(s))\|^2 \, ds < +\infty, \text{ and } \|M(y(s))\| = o\left(\frac{1}{s^2}\right) \text{ as } s \to +\infty.$$

It would be interesting to know if we can achieve the same convergence rate using time scaling and perturbation techniques.

5.2. Additively structured monotone inclusions.　Our interest now is to solve the structured monotone inclusion problem

$$0 \in A(x) + B(x),$$

where A is maximally monotone, and B is cocoercive with $(A+B)^{-1}(0) \neq \emptyset$. A favorable situation occurs when one can compute the resolvent operator of A

$$J_{\mu A} = (I + \mu A)^{-1}, \quad \mu > 0.$$

In this case, we can develop a strategy parallel to the one which consists in replacing a maximally monotone operator by its Yosida approximation. Indeed, given $\mu > 0$, we have

$$(A + B)(x) \ni 0 \iff x - J_{\mu A}(x - \mu B(x)) = 0 \iff M_{A,B,\mu}(x) = 0, \tag{132}$$

where $M_{A,B,\mu} : \mathcal{H} \to \mathcal{H}$ is the single-valued operator defined by

$$M_{A,B,\mu}(x) = \frac{1}{\mu} \left(x - J_{\mu A}(x - \mu B(x)) \right). \tag{133}$$

The operator $M_{A,B,\mu}$ is closely tied to the well-known forward-backward fixed point operator. Moreover, when $B = 0$, $M_{A,B,\mu} = \frac{1}{\mu} \left(I - J_{\mu A} \right)$ which is nothing but the Yosida regularization of A with index μ. As a remarkable property, for the paremeter μ properly set, the operator $M_{A,B,\mu}$ is cocoercive. This is made precise in the following result, whose proof relies on the relation between cocoercivity and averaged property.

PROPOSITION 1.   ([36, Proposition 4.4]) *Let* $A : \mathcal{H} \to 2^{\mathcal{H}}$ *be a general maximally monotone operator, and* $B : \mathcal{H} \to \mathcal{H}$ *a monotone operator which is* λ*-cocoercive. Assume that* $\mu \in]0, 2\lambda[$. *Then* $M_{A,B,\mu}$ *is* ρ*-cocoercive with*

$$\rho = \mu \left(1 - \frac{\mu}{4\lambda} \right).$$

Therefore, we can develop a strategy parallel to that developed in [21], which consists in replacing a maximally monotone operator by its Yosida approximation. This leads to the inertial dynamics

$$\ddot{y}(s) + \frac{\alpha}{s} \dot{y}(s) + \frac{s}{\alpha + 1} \frac{d}{ds} \left(M_{A,B,\mu}(y(s)) \right) + M_{A,B,\mu}(y(s)) = 0. \tag{134}$$

It is immediate to apply Theorem 13 to (134). This open new perspectives for forward-backward accelerated regularized Newton methods.

Let us illustrate our results by considering the following composite optimization problem

$$\min_{y \in \mathbb{R}^n} \left\{ f(y) := \frac{1}{2} \|Ay - b\|^2 + g(y) \right\},$$

where A is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function which acts as a regularizer. This class of problems covers many interesting situations arising in the signal and image processing such as the LASSO problem. We are in the setting of (133) with

$$M_{A,B,\mu}(y) := \frac{1}{\mu} \left(y - \mathrm{prox}_{\mu g}(y + \mu A^*(b - Ay)) \right),$$

42

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

where $0 < \mu < \frac{1}{\|A\|^2}$. Following [9], we can also interpret our problem as minimizing the Moreau envelope of f computed for an ad hoc metric. Precisely, if $W := \frac{1}{\mu}\mathbb{I} - A^*A$ and

$$f_W(y) := \arg\min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2}\|y - u\|_W^2 \right\},$$

then $\nabla f_W(y) = M_{A,B,\mu}(y)$ and $\inf_{\mathcal{H}} f = \inf_{\mathcal{H}} f_W$. So, we can use this example to illustrate also the convergence properties of the dynamic with explicit Hessian damping considered in Subsection 3.1. Thus, we can expect a convergence rate of $1/s^2$ for $f(y_W(s)) - \inf_{\mathcal{H}} f$, where

$$y_W(s) := \text{prox}_{\mu g}(y(s) + \mu A^*(b - Ay(s))).$$

We conduct the experiment for a randomly generated matrix A and by taking g to be the ℓ_1 norm with various choices for $\alpha \in \{1.001, 2, 3, 5\}$. Notice that we do not need need to solve (134) directly but rather exploit the fact that it can be rewritten as the first order equation

$$\dot{y}(s) + \frac{s}{\alpha + 1}M_{A,B,\mu}(y(s)) = \frac{s_0^\alpha}{s^\alpha}\left(y_1 + \frac{s_0}{\alpha + 1}y_0\right),$$

as shown previously. We then use the Runge-Kutta method to solve this equation numerically. Below we plot for the different choices of α the values of f at $y_W(s)$ and of the norm of the operator at $y(s)$, respectively.
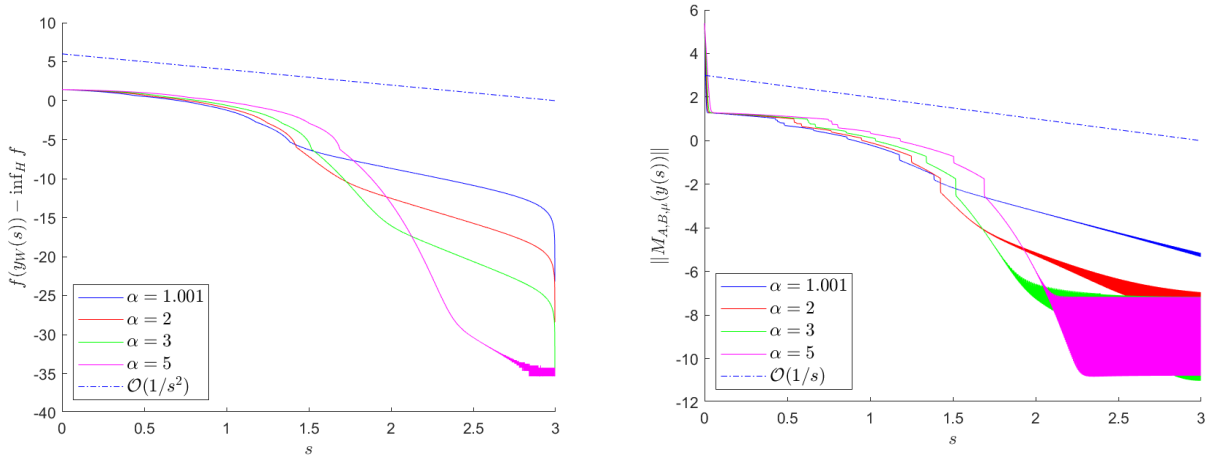

FIGURE 1. Convergence of the function values and of the norm of the operator values in logarithmic scale.

6. Conclusion and perspectives.

Our study has shed new light on the acceleration of gradient methods for general convex differentiable optimization. Based on a dynamical approach, and using relatively elementary tools, namely time scaling and time averaging, we were able to

introduce a large class of damped inertial dynamics with proven fast convergence rates. The method is very flexible and can be adapted to various initial start-up dynamics. We paid particular attention to the steepest continuous descent and the regularized Newton method. Our approach confirms the important role played by the Hessian driven damping role in these questions. Our results open the door to a systematic study of the corresponding fast proximal and gradient methods. Importantly, our approach also works when we take a cocoercive operator instead of a gradient operator. Applications to inertial forward-backward methods, ADMM, can therefore be expected, to cite some of the most popular algorithms for convex structured optimization. Application to stochastic differential equations can be considered as well.

Appendix A: Auxiliary results. Opial's lemma in continuous form is a basic ingredient of the asymptotic analysis of dynamical systems.

LEMMA 1. (Opial) *Let S be a nonempty subset of $\mathcal{H}$ and let $x\colon [t_0,+\infty[\to\mathcal{H}$. Assume that*
 (i) *for every $z\in S$, $\lim_{t\to\infty}\|x(t)-z\|$ exists;*
 (ii) *every weak sequential limit point of $x(t)$, as $t\to\infty$, belongs to S.*
Then $x(t)$ converges weakly as $t\to+\infty$, and its limit belongs to S.

In the analysis of the perturbed dynamical systems Gronwall-Bellman's lemma (see [31, Lemme A.5]) plays an important role.

LEMMA 2. *Let $T\geq\delta\geq 0$ and $g\colon [\delta,T]\to\mathbb{R}^+$ be integrable, and let $c\geq 0$. Suppose that $h\colon [\delta,T]\to\mathbb{R}$ is continuous and*
$$\frac{1}{2}h^2(t)\leq\frac{1}{2}c^2+\int_\delta^t h(t)\,g(t)\,dt$$
for every $t\in[\delta,T]$. Then $|h(t)|\leq c+\int_\delta^t g(u)\,du$ for every $t\in[\delta,T]$.

LEMMA 3. *Let $\delta\geq 0$ and $h\colon [\delta,+\infty]\to\mathbb{R}^+$ be a nonincreasing function belonging to $\mathbb{L}^1([\delta,+\infty[\,;\mathbb{R})$. It holds $\lim_{t\to+\infty}th(t)=0$.*

Proof. The nonincreasing property of h implies that $\frac{d}{dt}(th(t))=h(t)+t\dot{h}(t)\leq h(t)$ for every $t\geq t_0$. Since $h\in\mathbb{L}^1([\delta,+\infty])$, the result follows from [1, Lemma 5.2]. Q.E.D.

LEMMA 4. *Let $\delta\geq 0$ and $a\colon [\delta,+\infty[\to\mathbb{R}^+$ be an integrable function such that $\lim_{s\to+\infty}a(s)=0$. Then $\lim_{s\to+\infty}A(s)=0$, where for $p>0$*
$$A(s)=\frac{1}{s^p}\int_{s_0}^s u^{p-1}a(u)\,du.$$

44

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

Proof. Given $\epsilon > 0$, let T_ϵ such that $s_0 < T_\epsilon$ and $a(u) \leq \epsilon$ for $t \geq T_\epsilon$. For $s > T_\epsilon$, we have

$$A(s) = \frac{1}{s^p} \int_{s_0}^{T_\epsilon} u^{p-1} a(u) du + \frac{1}{s^p} \int_{T_\epsilon}^{s} u^{p-1} a(u) du \leq \frac{1}{s^p} \int_{s_0}^{T_\epsilon} u^{p-1} a(u) du + \epsilon.$$

Letting s converge to $+\infty$, it yields

$$\limsup_{s \to +\infty} A(s) \leq \epsilon.$$

This being true for any $\epsilon > 0$, we infer that $\lim_{s \to +\infty} A(s) = 0$, which gives the claim.   Q.E.D.

Appendix B: Missing proofs.

Proof of Theorem 1.   For sake of completeness, let us recall some of the arguments used in the asymptotic analysis of the perturbed steepest descent system when the perturbation g satisfies (7). Given $z_* \in S$, let $T > t_0$ be fixed and for every $T \geq t \geq t_0$ consider

$$\mathcal{E}_T(t) := t(f(z(t)) - \inf_{\mathcal{H}} f) + \frac{1}{2} \|z(t) - z_*\|^2 + \int_t^T \langle z(u) - z_* + \tau \dot{z}(u), g(u) \rangle du.$$

Differentiating $\mathcal{E}_T$ gives for every $T \geq t \geq t_0$

$$\frac{d}{dt}\mathcal{E}_T(t) = f(z(t)) - \inf_{\mathcal{H}} f + t\langle \nabla f(z(t)), \dot{z}(t) \rangle + \langle z(t) - z_*, \dot{z}(t) - g(t) \rangle - t\langle \dot{z}(t), g(t) \rangle$$

$$= f(z(t)) - \inf_{\mathcal{H}} f - t\|\dot{z}(t)\|^2 - \langle z(t) - z_*, \nabla f(z(t)) \rangle$$

$$\leq -t\|\dot{z}(t)\|^2,$$

where the second equality comes from (6) and the last inequality follows from the convexity of f. By integration from t_0 to t and by denoting $C_3 := t_0(f(z(t_0)) - \inf_{\mathcal{H}} f) + \frac{1}{2} \|z(t_o) - z_*\|^2 \geq 0$ we deduce that

$$t(f(z(t)) - \inf_{\mathcal{H}} f) + \frac{1}{2} \|z(t) - z_*\|^2$$

$$\leq C_3 + \int_{t_0}^t \langle z(u) - z_*, g(u) \rangle du + \int_{t_0}^t \tau \langle \dot{z}(u), g(u) \rangle du - \int_{t_0}^t \tau \|\dot{z}(u)\|^2 du$$

$$\leq C_3 + \int_{t_0}^t \|z(u) - z_*\| \|g(u)\| du + \frac{1}{2} \int_{t_0}^{+\infty} \tau \|g(u)\|^2 du - \frac{1}{2} \int_{t_0}^t \tau \|\dot{z}(u)\|^2 du. \tag{135}$$

We obtain that there exists $C_4 \geq 0$ such that the following estimate is true for every $t \geq t_0$

$$\frac{1}{2} \|z(t) - z_*\|^2 \leq C_4 + \int_{t_0}^t \|z(u) - z_*\| \|g(u)\| du.$$

According to Gronwall's lemma (see Lemma 2 in Appendix A), we conclude that for every $t \geq t_0$

$$\|z(t) - z_*\| \leq \sqrt{2C_4} + \int_{t_0}^{t} \|g(u)\| \, du \leq \sqrt{2C_4} + \int_{t_0}^{+\infty} \|g(u)\| \, du < +\infty.$$

The trajectory is therefore bounded. Using this property and condition (7) allow us to assert from (135) that there exists $C_5 \geq 0$ such that for every $t \geq t_0$

$$t \left(f(z(t)) - \inf_{\mathcal{H}} f \right) + \frac{1}{2} \|z(t) - z_*\|^2 + \frac{1}{2} \int_{t_0}^{t} \tau \|\dot{z}(u)\|^2 \, du \leq C_5. \tag{136}$$

Since $t \left(f(z(t)) - \inf_{\mathcal{H}} f \right) + \frac{1}{2} \|z(t) - z_*\|^2 \geq 0$ for every $t \geq t_0$, letting t tend to $+\infty$ in (136) we deduce that

$$\int_{t_0}^{+\infty} t \|\dot{z}(t)\|^2 \, dt < +\infty.$$

According to the constitutive equation (6) we get item (ii)

$$\int_{t_0}^{+\infty} t \|\nabla f(z(t))\|^2 \, dt \leq 2 \int_{t_0}^{+\infty} t \|\dot{z}(t)\|^2 \, dt + 2 \int_{t_0}^{+\infty} t \|g(t)\|^2 \, dt < +\infty.$$

Differentiate the anchor function, it yields for every $t \geq t_0$

$$\begin{aligned}
\frac{d}{dt} \left(\frac{1}{2} \|z(t) - z_*\|^2 \right) &= \langle z(t) - z_*, \dot{z}(t) \rangle \\
&= -\langle z(t) - z_*, \nabla f(z(t)) \rangle - \langle z(t) - z_*, g(t) \rangle \\
&\leq -\left(f(z(t)) - \inf_{\mathcal{H}} f \right) + \sup_{u \geq t_0} \|z(u) - z_*\| \cdot \|g(t)\| \tag{137} \\
&\leq \sup_{u \geq t_0} \|z(u) - z_*\| \cdot \|g(t)\|. \tag{138}
\end{aligned}$$

Recall that the trajectory $z(\cdot)$ is bounded. Now let us show that in fact

$$\lim_{t \to +\infty} t \left(f(z(t)) - \inf_{\mathcal{H}} f \right) = 0,$$

meaning that the convergence rate of $f(z(t)) - \inf_{\mathcal{H}} f$ is actually $o(1/t)$. To this end we integrate (137) from t_0 to $t > t_0$ and let then t converge to $+\infty$. This yields (iii)

$$\int_{t_0}^{+\infty} \left(f(z(t)) - \inf_{\mathcal{H}} f \right) dt = \int_{t_0}^{+\infty} \frac{1}{t} t \left(f(z(t)) - \inf_{\mathcal{H}} f \right) dt < +\infty \tag{139}$$

46

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

and thus $\liminf_{t\to+\infty} t\,(f\,(z\,(t)) - \inf_{\mathcal{H}} f) = 0$. It remains to show that the limit exists. To this end we compute the time derivative of $t\,(f\,(z\,(t)) - \inf_{\mathcal{H}} f)$ for every $t \geq t_0$

$$
\begin{aligned}
\frac{d}{dt}\,(t\,(f\,(z\,(t)) - \inf_{\mathcal{H}} f)) &= f\,(z\,(t)) - \inf_{\mathcal{H}} f + t\,\langle \nabla f\,(z\,(t)), \dot{z}\,(t)\rangle \\
&= f\,(z\,(t)) - \inf_{\mathcal{H}} f - t\,\|\dot{z}\,(t)\|^2 - t\,\langle g\,(t), \dot{z}\,(t)\rangle \\
&\leq f\,(z\,(t)) - \inf_{\mathcal{H}} f + \frac{1}{4}t\,\|g\,(t)\|^2
\end{aligned}
$$

and apply once again [1, Lemma 5.1]. Statement (iv) follows from assumption (7) and (139). According to assumption (7), we deduce that the right hand side of (138) belongs to $\mathbb{L}^1\,([t_0, +\infty[\,; \mathbb{R})$. Therefore, from [1, Lemma 5.1] we obtain that $\lim_{t\to+\infty} \|z\,(t) - z_*\|^2 \in \mathbb{R}$ exists, and so $\lim_{t\to+\infty} \|z\,(t) - z_*\| \in \mathbb{R}$ does. In other words, the first condition of Opial's lemma (see Lemma 1 in Appendix A) is fulfilled. Furthermore, since $\lim_{t\to+\infty} f\,(z\,(t)) = \inf_{\mathcal{H}} f$ and f is convex and weakly lower semicontinuous, the second condition of Opial's lemma is also fulfilled. This gives the weak convergence of the trajectory $z(t)$ as $t \to +\infty$ to an element in $S = \arg\min f$.

The proof of (vi) follows similarly. First, the assertion (ii) gives $\liminf_{t\to+\infty} t^2\,\|\nabla f\,(z\,(t))\|^2 = 0$. Let L be the Lipschitz constant of ∇f on a ball containing the trajectory $z(\cdot)$. We have for almost every $t \geq t_0$

$$
\begin{aligned}
&\frac{d}{dt}\,\left(t^2\,\|\nabla f\,(z\,(t))\|^2\right) \\
&= 2t\,\|\nabla f\,(z\,(t))\|^2 + t^2\,\left\langle \nabla f\,(z\,(t)), \frac{d}{dt}\nabla f\,(z\,(t))\right\rangle \\
&= 2t\,\|\nabla f\,(z\,(t))\|^2 - t^2\,\left\langle \dot{z}\,(t), \frac{d}{dt}\nabla f\,(z\,(t))\right\rangle + t^2\,\left\langle g\,(t), \frac{d}{dt}\nabla f\,(z\,(t))\right\rangle \\
&\leq 2t\,\|\nabla f\,(z\,(t))\|^2 - \frac{t^2}{L}\,\left\|\frac{d}{dt}\nabla f\,(z\,(t))\right\|^2 + 4Lt^2\,\|g\,(t)\|^2 + \frac{t^2}{L}\,\left\|\frac{d}{dt}\nabla f\,(z\,(t))\right\|^2 \\
&= 2t\,\|\nabla f\,(z\,(t))\|^2 + 4Lt^2\,\|g\,(t)\|^2.
\end{aligned}
$$

The convergence rate follows by applying once again [1, Lemma 5.1] and by using that $t \mapsto tg\,(t) \in \mathbb{L}^2\,([t_0, +\infty[\,; \mathcal{H})$. Q.E.D.

Proof of Theorem 5. Given $z_* \in S$, we define the following function for every $t \geq t_0$, which is nonnegative due to the convexity of f and the fact that $v\,(t) = \nabla f\,(z\,(t))$

$$
\phi\,(t) := [f\,(z_*) - f\,(z\,(t)) - \langle v\,(t), z_* - z\,(t)\rangle] + \frac{\lambda}{2}\,\|z_* - z\,(t)\|^2.
$$

According to (62) and the convexity of f we have for every $t \geq t_0$

$$\frac{d}{dt}\phi(t) = -\langle v(t), \dot{z}(t) \rangle - \langle \dot{v}(t), z_* - z(t) \rangle + \langle v(t), \dot{z}(t) \rangle - \lambda \langle z_* - z(t), \dot{z}(t) \rangle$$

$$= \langle v(t), z_* - z(t) \rangle \leq \inf_{\mathcal{H}} f - f(z(t)) \leq 0.$$

We integrate from t_0 to $t > t_0$ and let t tend to $+\infty$, and obtain (i).

The inequality above also implies that the function ϕ is nonincreasing on $[t_0, +\infty[$, from which we deduce that $\phi(t)$ converges as $t \to +\infty$, and that the trajectory $z(\cdot)$ is bounded. Let $L > 0$ be the Lipschitz constant of ∇f on a ball that contains the trajectory $z(\cdot)$. It follows from (62) that for every $t \geq t_0$

$$\frac{d}{dt}(f(z(t)) - \inf_{\mathcal{H}} f) = \langle v(t), \dot{z}(t) \rangle = -\lambda \|\dot{z}(t)\|^2 - \langle \dot{v}(t), \dot{z}(t) \rangle$$

$$= -\lambda \|\dot{z}(t)\|^2 - \left\langle \frac{d}{dt}(\nabla f(z(t))), \dot{z}(t) \right\rangle$$

$$\leq -\lambda \|\dot{z}(t)\|^2 - \frac{1}{L}\left\| \frac{d}{dt}(\nabla f(z(t))) \right\|^2 = -\lambda \|\dot{z}(t)\|^2 - \frac{1}{L}\|\dot{v}(t)\|^2 \leq 0.$$

Since $t \mapsto f(z(t)) - \inf_{\mathcal{H}} f$ is nonincreasing, by applying Lemma 3 in the Appendix we obtain (ii).

In addition, we derive that

$$\frac{d}{dt}(t(f(z(t)) - \inf_{\mathcal{H}} f)) = f(z(t)) - \inf_{\mathcal{H}} f + \frac{d}{dt}(f(z(t)) - \inf_{\mathcal{H}} f)$$

$$\leq f(z(t)) - \inf_{\mathcal{H}} f - \lambda t \|\dot{z}(t)\|^2 - \frac{1}{L}t\|\dot{v}(t)\|^2$$

Integrating again from t_0 to $t > t_0$ and letting t tend to $+\infty$, we obtain (iii). Indeed, for every $t \geq t_0$ we have $\|v(t)\|^2 \leq 2\lambda^2 \|\dot{z}(t)\|^2 + 2\|\dot{v}(t)\|^2$, therefore

$$\int_{t_0}^{+\infty} t\|v(t)\|^2\, dt = \int_{t_0}^{+\infty} t\|\nabla f(z(t))\|^2\, dt < +\infty. \tag{140}$$

Consequently, for item iv) we can use again Lemma 3. Precisely, observe that for every $t \geq t_0$

$$\frac{d}{dt}\left(\frac{1}{2}\|v(t)\|^2\right) = \langle v(t), \dot{v}(t) \rangle = -\lambda \langle \dot{z}(t), \dot{v}(t) \rangle - \|\dot{v}(t)\|^2 \leq 0,$$

48

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

which means $t \mapsto \frac{1}{2} \|v(t)\|^2$ is nonincreasing. The conclusion follows by using (140) and by noticing that

$$\|v(t)\|^2 = \|\lambda \dot{z}(t) + \dot{v}(t)\|^2 = \lambda^2 \|\dot{z}(t)\|^2 + \|\dot{v}(t)\|^2 + 2\lambda \langle \dot{z}(t), \dot{v}(t) \rangle \geq \lambda^2 \|\dot{z}(t)\|^2 + \|\dot{v}(t)\|^2.$$

Finally, let us recall that $\lim_{t \to +\infty} \phi(t) \in \mathbb{R}$ exists. Moreover, according to (ii), (iv), and the fact that the trajectory $z(\cdot)$ is bounded, we conclude that

$$\lim_{t \to +\infty} \frac{\lambda}{2} \|z_* - z(t)\|^2 = \lim_{t \to +\infty} \phi(t) \in \mathbb{R}.$$

This yields the first condition of Opial's lemma. Now let $\widehat{z}$ be a weak sequential cluster point of the trajectory. Since $v(t) = \nabla f(z(t))$ for every $t \geq t_0$, it follows from the sequential closedness property of ∇f in the weak $\times$ strong topology on $\mathcal{H} \times \mathcal{H}$ that $\widehat{z} \in S$, which proves that the second condition of Opial's lemma is also satisfied. This completes the proof.   Q.E.D.

Proof of Theorem 12.   For the sake of completeness, we briefly sketch the Lyapunov analysis of (127). Given $z_* \in S$, let us first fix $T > t_0$, and for every $T \geq t \geq t_0$ consider

$$\mathcal{E}_T(t) := \frac{1}{2} \|z(t) - z_*\|^2 + \int_t^T \langle z(u) - z_*, g(u) \rangle \, du.$$

For every $T \geq t \geq t_0$ it holds

$$\frac{d}{dt} \mathcal{E}_T(t) = \langle z(t) - z_*, \dot{z}(t) \rangle - \langle z(t) - z_*, g(t) \rangle = - \langle z(t) - z_*, M(z(t)) \rangle.$$

By integration from t_0 to t, we get for $C_6 := \frac{1}{2} \|z(t_0) - z_*\|^2$

$$\frac{1}{2} \|z(t) - z_*\|^2 + \int_{t_0}^t \langle z(t) - z_*, M(z(t)) \rangle \, du \leq C_6 + \int_{t_0}^t \langle z(u) - z_*, g(u) \rangle \, du$$

$$\leq C_6 + \int_{t_0}^t \|z(u) - z_*\| \|g(u)\| \, du. \qquad (141)$$

By Gronwall's lemma (see Lemma 2 in Appendix A), we conclude that for every $t \geq t_0$

$$\|z(t) - z_*\| \leq \sqrt{2C_6} + \int_{t_0}^t \|g(u)\| \, du \leq \sqrt{2C_6} + \int_{t_0}^{+\infty} \|g(u)\| \, du < +\infty.$$

This gives the boundedness of the trajectory. Using this assertion in (141), we deduce that for every $t \geq t_0$

$$
\int_{t_0}^{t} \langle z(u) - z_*, M(z(u)) \rangle \, du \leq C_6 + \sup_{t \geq t_0} \|z(u) - z_*\| \int_{t_0}^{t} \|g(u)\| \, du
$$

$$
\leq C_6 + \sup_{t \geq t_0} \|z(u) - z_*\| \int_{t_0}^{+\infty} \|g(u)\| \, du < +\infty.
$$

According to the ρ-cocoercivity of M, and by letting t go to infinity, we obtain

$$
\rho \int_{t_0}^{+\infty} \|M(z(t))\|^2 \, dt \leq \int_{t_0}^{+\infty} \langle z(t) - z_*, M(z(t)) \rangle \, dt < +\infty. \tag{142}
$$

Therefore, $\liminf_{t \to +\infty} t \|M(z(t))\|^2 = 0$. Let us now compute the time derivative for every $t \geq t_0$

$$
\frac{d}{dt} \left(\frac{1}{2} t \|M(z(t))\|^2 \right) = \frac{1}{2} \|M(z(t))\|^2 + t \left\langle M(z(t)), \frac{d}{dt}(M(z(t))) \right\rangle
$$

$$
= \frac{1}{2} \|M(z(t))\|^2 - t \left\langle \dot{z}(t), \frac{d}{dt}(M(z(t))) \right\rangle + t \left\langle g(t), \frac{d}{dt}(M(z(t))) \right\rangle.
$$

Using the cocoercivity of M and the Cauchy-Schwarz inequality, we deduce that

$$
- t \left\langle \dot{z}(t), \frac{d}{dt}(M(z(t))) \right\rangle + t \left\langle g(t), \frac{d}{dt}(M(z(t))) \right\rangle
$$

$$
\leq -\rho t \left\| \frac{d}{dt}(M(z(t))) \right\|^2 + t \|g(t)\| \left\| \frac{d}{dt}(M(z(t))) \right\|
$$

$$
\leq \frac{1}{4\rho} t \|g(t)\|^2.
$$

The right hand side of the above inequality belongs to $\mathbb{L}^1([t_0, +\infty[\,;\mathbb{R})$ thanks to (128) and (142). According to [1, Lemma 5.1] we have that $\lim_{t \to +\infty} t \|M(z(t))\|^2 \in \mathbb{R}$ exists, and therefore it must be equal to 0

$$
\|M(z(t))\| = o\left(\frac{1}{\sqrt{t}} \right) \text{ as } t \to +\infty.
$$

This also means that $\lim_{t \to +\infty} M(z(t)) = 0$. According to the sequential closedness property of the graph of M in the weak $\times$ strong topology on $\mathcal{H} \times \mathcal{H}$, every weak limit point of the trajectory z

50

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

belongs to S. It remains to verify the first condition of Opial's lemma. Given $z_* \in S$, observe that for every $t \geq t_0$ it holds

$$\frac{d}{dt}\left(\frac{1}{2}\|z(t) - z_*\|^2\right) = \langle z(t) - z_*, \dot{z}(t)\rangle = -\langle z(t) - z_*, M(z(t))\rangle + \langle z(t) - z_*, g(t)\rangle$$

$$\leq \sup_{t \geq t_0}\|z(t) - z_*\| \cdot \|g(t)\|.$$

This shows that the limit $\lim_{t \to +\infty}\|z(t) - z_*\| \in \mathbb{R}$ thanks to (128) and [1, Lemma 5.1].   Q.E.D.

Acknowledgments

References

[1] Abbas B, Attouch H, Svaiter B (2014) Newton-like dynamics and forward–backward methods for structured monotone inclusions in hilbert spaces. *J. Optim. Theory Appl.* 161(2):331–360, URL http://dx.doi.org/10.1007/s10957-013-0414-5.

[2] Alecsa C, László S, Pinta T (2021) An extension of the second order dynamical system that models Nesterov's convex gradient method. *Appl. Math. Optim.* 84:1687–1716, URL http://dx.doi.org/10.1007/s00245-020-09692-1.

[3] Attouch H, Balhag A, Chbani Z, Riahi H (2022) Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. *Evol. Equ. Control Theory* 11(2):487–514, URL http://dx.doi.org/10.3934/eect.2021010.

[4] Attouch H, Boţ RI, Csetnek ER (2023) Fast optimization via inertial dynamics with closed-loop damping. *J. Eur. Math. Soc.* 5(25):1985–2056, URL http://dx.doi.org/10.4171/JEMS/1231.

[5] Attouch H, Buttazzo G, Michaille G (2014) *Variational Analysis in Sobolev and BV spaces. Applications to PDE's and Optimization*. MOS/SIAM Series on Optimization, MO 17 (Philadelphia: Society for Industrial and Applied Mathematics (SIAM)), 2nd edition.

[6] Attouch H, Cabot A (2017) Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *J. Differ. Equ.* 263(9):5412–5458, URL http://dx.doi.org/10.1016/j.jde.2017.06.024.

[7] Attouch H, Cabot A (2018) Convergence rates of inertial forward-backward algorithms. *SIAM J. Optim.* 28(1):849–874, URL http://dx.doi.org/10.1137/17M1114739.

Attouch, Boț and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

51

[8] Attouch H, Cabot A (2020) Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Math. Program.* 184:243–287, URL http://dx.doi.org/10.1007/s10957-013-0414-5.

[9] Attouch H, Chbani Z, Fadili J, Riahi H (2022) First-order algorithms via inertial systems with hessian driven damping. *Math. Program.* 193:113–155, URL http://dx.doi.org/10.1007/s10107-020-01591-1.

[10] Attouch H, Chbani Z, Fadili J, Riahi H (2023) Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping. *Optimization* 72(5):1199–1238, URL http://dx.doi.org/10.1080/02331934.2021.2009828.

[11] Attouch H, Chbani Z, Peypouquet J, Redont P (2018) Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Math. Program. Ser. B* 168:123–175, URL http://dx.doi.org/10.1007/s10107-016-0992-8.

[12] Attouch H, Chbani Z, Riahi H (2019) Fast proximal methods via time scaling of damped inertial dynamics. *SIAM J. Optim.* 29(3):2227–2256, URL http://dx.doi.org/10.1137/18M1230207.

[13] Attouch H, Chbani Z, Riahi H (2022) Fast convex optimization via a third-order in time evolution equation. *Optimization* 71(5):1275–1304, URL http://dx.doi.org/10.1080/02331934.2020.1764953.

[14] Attouch H, Chbani Z, Riahi H (2022) Fast convex optimization via a third-order in time evolution equation : TOGES-V an improved version of toges. *Optimization* URL http://dx.doi.org/10.1080/02331934.2022.2119084.

[15] Attouch H, Czarnecki MO (2010) Asymptotic behavior of coupled dynamical systems with multiscale aspects. *J. Diff. Equ.* 248(6):1315–1344, URL http://dx.doi.org/10.1007/s10957-013-0414-5.

[16] Attouch H, Fadili J (2022) From the Ravine method to the Nesterov method and vice versa: A dynamical system perspective. *SIAM J. Optim.* 32(3):2074–2101, URL http://dx.doi.org/10.1137/22M1474357.

[17] Attouch H, Fadili J, Kungurtsev V (2023) On the effect of perturbations, errors in first-order optimization methods with inertia and hessian driven damping. *Evol. Equ. Control Theory* 1(12):71–117, URL http://dx.doi.org/10.3934/eect.2022022.

[18] Attouch H, Goudou X, Redont P (2000) The heavy ball with friction method. the continuous dynamical system, global exploration of the local minima of a real-valued function by asymptotical analysis of a dissipative dynamical system. *Commun. Contemp. Math.* 2(1):1–34, URL http://dx.doi.org/10.1142/S0219199700000025.

[19] Attouch H, Marques-Alves M, Svaiter BF (2016) A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity $O(1/n^2)$. *J. Convex Anal.* 23(1):139–180.

[20] Attouch H, Peypouquet J (2016) The rate of convergence of nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM J. Optim.* 26(3):1824–1834, URL http://dx.doi.org/10.1137/15M1046095.

52

Attouch, Boţ and Nguyen: *Fast Convex Optimization via Time Scale and Averaging of the Steepest Descent*
Article submitted to *Mathematics of Operations Research*

[21] Attouch H, Peypouquet J (2019) Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators. *Math. Program.* 174(1):391–432, URL http://dx.doi.org/10.1007/s10107-018-1252-x.

[22] Attouch H, Peypouquet J, Redont P (2016) Fast convex optimization via inertial dynamics with hessian driven damping. *J. Diff. Equ.* 261(10):5734–5783, URL http://dx.doi.org/10.1016/j.jde.2016.08.020.

[23] Attouch H, Redont P, Svaiter BF (2013) Global convergence of a closed-loop regularized Newton method for solving monotone inclusions in Hilbert spaces. *J. Optim. Theory Appl.* 157(3):624–650, URL http://dx.doi.org/10.1007/s10957-012-0222-3.

[24] Attouch H, Svaiter BF (2011) A continuous dynamical Newton-like approach to solving monotone inclusions. *SIAM J. Control Optim.* 49(2):574–598, URL http://dx.doi.org/10.1137/100784114.

[25] Bauschke H, Combettes PL (2017) *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CSM Books in Mathematics (New York: Springer), 2nd edition.

[26] Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2(1):183–202, URL http://dx.doi.org/10.1137/080716542.

[27] Boţ RI, Csetnek ER (2017) A dynamical system associated with the fixed points set of a nonexpansive operator. *J. Dyn. Diff. Equ.* 29:155–168, URL http://dx.doi.org/10.1007/s10884-015-9438-x.

[28] Boţ RI, Csetnek ER, László S (2021) Tikhonov regularization of a second order dynamical system with Hessian driven damping. *Math. Program.* 189:151–186, URL http://dx.doi.org/10.1007/s10957-013-0414-5.

[29] Boţ RI, Csetnek ER, Nguyen DK (2023) Fast optimistic gradient descent ascent (OGDA) method in continuous and discrete time. *Found. Comp. Math.* URL http://dx.doi.org/10.1007/s10208-023-09636-5.

[30] Boţ RI, Nguyen DK (2021) Improved convergence rates and trajectory convergence for primal-dual dynamical systems with vanishing damping. *J. Differ. Equ.* 303:369–406, URL http://dx.doi.org/10.1016/j.jde.2021.09.021.

[31] Brézis H (1972) *Opérateurs Maximaux Monotones dans les espaces de Hilbert et Équations d Évolution*. Lecture Notes 5 (Amsterdam: North Holland).

[32] Brézis H (1983) *Analyse Fonctionnelle, Théorie et Applications* (Amsterdam: North Holland).

[33] Cabot A, Engler H, Gadat S (2009) On the long time behavior of second order differential equations with asymptotically small dissipation. *Trans. Amer. Math. Soc.* 361(11):5983–6017, URL http://www.jstor.org/stable/40302945.

[34] Castera C, Bolte J, Févotte C, Pauwels E (2021) Newton-like dynamics and forward–backward methods for structured monotone inclusions in hilbert spaces. *J. Mach. Learn. Res.* 22(134):1–31, URL http://jmlr.org/papers/v22/19-1024.html.

[35] Chambolle A, Dossal C (2015) On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm. *J. Optim. Theory Appl.* 166:968–982, URL http://dx.doi.org/10.1007/s10957-015-0746-4.

[36] Combettes P, Yamada I (2015) Compositions and convex combinations of averaged nonexpansive operators. *J. Math. Anal. Appl.* 425(1):55–70, URL http://dx.doi.org/10.1016/j.jmaa.2014.11.044.

[37] Gelfand I, Tsetlin M (1963) Printszip nelokalnogo poiska v sistemah avtomatich. *Optimizatsii, Dokl. AN SSSR* 137:295–298, URL http://dx.doi.org/10.1007/s10957-013-0414-5.

[38] Lin T, Jordan M (2023) On monotone inclusions, acceleration and closed-loop control. *Mathematics of Operations Research* 48(4):1811–2382, URL http://dx.doi.org/10.1287/moor.2022.1343.

[39] May R (2017) Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. *Turkish Journal of Mathematics* 41(3):681–685.

[40] Nesterov Y (1983) A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady* 27:372–376.

[41] Nesterov Y (2004) *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization (Boston: Kluwer Academic Publishers).

[42] Nesterov Y (2005) Smooth minimization of non-smooth functions. *Math. Program.* 103:127–152.

[43] Polyak B (1964) Some methods of speeding up the convergence of iteration methods. *U.S.S.R. Comput. Math. Math. Phys.* 4:1–17, URL http://dx.doi.org/10.1016/0041-5553(64)90137-5.

[44] Polyak B (1987) *Introduction to Optimization*. Optimization Software (New York).

[45] Poveda JI, Li N (2021) Robust hybrid zero-order optimization algorithms with acceleration via averaging in time. *Automatica* 123:109361, URL http://dx.doi.org/10.1016/j.automatica.2020.109361.

[46] Scieur D, D'Aspremont A, Bach F (2020) Regularized nonlinear acceleration. *Math. Program.* 179:47–83, URL http://dx.doi.org/10.1007/s10107-018-1319-8.

[47] Shi B, Du SS, Jordan MI, Su WJ (2022) Understanding the acceleration phenomenon via high-resolution differential equations. *Math. Program.* 195:79–148, URL http://dx.doi.org/10.1007/s10107-021-01681-8.

[48] Su WJ, Boyd S, Candès EJ (2016) A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res.* 17(153):1–43, URL http://jmlr.org/papers/v17/15-084.html.