



**HAL**  
open science

# Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions

Vassilis Apidopoulos, Jean-François Aujol, Charles H Dossal, Aude Rondepierre

► **To cite this version:**

Vassilis Apidopoulos, Jean-François Aujol, Charles H Dossal, Aude Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 2020, 10.1007/s10107-020-01476-3 . hal-01965095v3

**HAL Id: hal-01965095**

**<https://hal.science/hal-01965095v3>**

Submitted on 19 Sep 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions

Vassilis Apidopoulos<sup>1</sup>   Jean-François Aujol<sup>1</sup>   Charles Dossal<sup>2</sup>   Aude Rondepierre<sup>2,3</sup>

<sup>1</sup>Université de Bordeaux, IMB, UMR 5251, F-33400 Talence, France.  
CNRS, IMB, UMR 5251, F-33400 Talence, France.

<sup>2</sup>IMT, Université de Toulouse, INSA Toulouse, France.

<sup>3</sup>LAAS, Université de Toulouse, CNRS, Toulouse, France.

{vasileios.apidopoulos/jean-francois.ujol}@math.u-bordeaux.fr  
{charles.dossal/aude.rondepierre}@insa-toulouse.fr

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Geometry of convex functions around their minimizers</b>	<b>3</b>
<b>3</b>	<b>Main results and contributions</b>	<b>5</b>
3.1	Main results . . . . .	5
3.2	Extension to the non-differential setting . . . . .	11
<b>4</b>	<b>Asymptotic analysis</b>	<b>12</b>
4.1	Schema of proofs . . . . .	12
4.2	The Lyapunov energy . . . . .	13
4.3	Proofs . . . . .	14
4.3.1	Proof of Theorem 3.1 . . . . .	14
4.3.2	Proof of Theorem 3.2 . . . . .	17
<b>A</b>	<b>General Lemmas</b>	<b>21</b>
<b>B</b>	<b>Proofs of Lemmas in Sections 4.2 and 4.3</b>	<b>24</b>

### Abstract

In this paper we study the convergence properties of the Nesterov’s family of inertial schemes which is a specific case of inertial Gradient Descent algorithm in the context of a smooth convex minimization problem, under some additional hypotheses on the local geometry of the objective function  $F$ , such as the growth (or Łojasiewicz) condition. In particular we study the different convergence rates for the objective function and the local variation, depending on these geometric conditions. In this setting we can give optimal convergence rates for this Nesterov scheme. Our analysis shows that there are some situations when Nesterov’s family of inertial schemes is asymptotically less efficient than the gradient descent (e.g. in the case when the objective function is quadratic).

**Keywords:** Smooth optimization, convex optimization, inertial gradient descent algorithm, Nesterov acceleration, growth condition, Łojasiewicz condition, rate of convergence

## 1 Introduction

Let  $N \in \mathbb{N}^*$ . We are interested in the following minimization problem:

$$\min_{x \in \mathbb{R}^N} F(x) \tag{1.1}$$

where  $F : \mathbb{R}^N \rightarrow \mathbb{R}$ , is a convex function in  $\mathcal{C}^{1,1}(\mathbb{R}^N)$  with  $L$ -Lipschitz gradient such that  $\arg \min F \neq \emptyset$ . In this setting various algorithms have been proposed in order to solve numerically the minimization problem (1.1) such as the classical gradient descent algorithm and its variants.

Starting from a point  $x_0 \in \mathbb{R}^N$ , the classical Gradient Descent algorithm (1.2) with a fixed step-size  $\gamma > 0$  reads:

$$x_{n+1} = T_\gamma(x_n) := x_n - \gamma \nabla F(x_n). \quad (1.2)$$

Algorithm (1.2) is a descent scheme (i.e.  $F(x_{n+1}) \leq F(x_n)$ ) provided that  $0 < \gamma < \frac{2}{L}$ , and without any further hypothesis made on the function  $F$ , it provides a sequence that converges (weakly in an infinite-dimensional Hilbert space) to a minimizer, as also convergence rates for the values  $F(x_n) - F(x^*)$  of order  $o(n^{-1})$ . In addition this order has been proved to be optimal (see for example [33] and [22]).

In order to "accelerate" this convergence rate, in the seminal work of Nesterov [36], the author proposed an inertial version of Gradient descent algorithm with a suitable momentum term, i.e.

$$\begin{aligned} \theta_{n+1} &= \frac{1 + \sqrt{4\theta_n^2 + 1}}{2}, & \text{with } \theta_0 &= 1, \\ y_n &= x_n + \frac{\theta_n - 1}{\theta_{n+1}}(x_n - x_{n-1}), \\ x_{n+1} &= y_n - \gamma \nabla F(y_n). \end{aligned} \quad (1.3)$$

Algorithm (1.3) accelerates the convergence rate of the values  $F(x_n) - F(x^*)$ , in the sense that it is of order  $O(n^{-2})$  asymptotically. Several extensions have been made also in the non-differentiable setting (proximal-splitting-methods), see for example [23] and [11], as also with different choices for the momentum parameter (i.e. the term  $\frac{\theta_n - 1}{\theta_{n+1}}$  in (1.3)).

In this paper, we consider a special choice of the momentum parameter/over-relaxation term of the Inertial Gradient Descent (1.5). This inertial scheme is described hereafter and depends of a sequence  $(\alpha_n)_{n \geq 1}$ : for any  $0 < \gamma \leq \frac{1}{L}$ , we set  $x_0 = x_1 \in \mathbb{R}^N$  and for all  $n \geq 1$ :

$$\begin{aligned} y_n &= x_n + \alpha_n(x_n - x_{n-1}) \\ x_{n+1} &= T_\gamma(y_n) := y_n - \gamma \nabla F(y_n). \end{aligned} \quad (1.4)$$

$\alpha_n \in [0, 1]$  may be a constant parameter or it may depend on the iteration number. In this paper we focus on the Nesterov's family of algorithms, that is the specific choice  $\alpha_n = \frac{n}{n+b}$ , with  $b > 0$ . Note that this term is the same as the one chosen in [17] in the non-smooth setting (see also [45],[1], [7] and [6]). The term "family of algorithms" refers to the dependence on the parameter  $b > 0$ . We therefore rewrite explicitly here what we call Nesterov's family of inertial algorithms or simply Nesterov (inertial) scheme in the rest of the paper:

$$\begin{aligned} y_n &= x_n + \frac{n}{n+b}(x_n - x_{n-1}) \\ x_{n+1} &= T_\gamma(y_n) := y_n - \gamma \nabla F(y_n). \end{aligned} \quad (1.5)$$

Notice that in (1.3), we have  $1 - \frac{\theta_n - 1}{\theta_{n+1}} \sim \frac{3}{n}$ . Without any further hypothesis on the function  $F$ , it was proven that if the parameter  $b$  satisfies  $b \geq 3$ , then the convergence rate of the objective function is of order of  $O(n^{-2})$  (see for example, [6], [45] and [17]). Another interesting issue of this choice is that if  $b > 3$  it can be proven that the iterates of (1.5) converge to a minimizer (see for example [17] and [6], and it can also be shown that the order of convergence rate of the objective function is actually  $o(n^{-2})$  (see [8]). Other recent studies of algorithm (1.5) include results for the case  $b \in (0, 3)$  (see for example [1] and [7]) which provide an order of  $O(n^{-\frac{2b}{3}})$ .

In this work we are interested in studying the convergence properties of the Nesterov scheme that is inertial gradient descent scheme (1.5) for solving the minimization problem (1.1), under some additional assumptions on the local geometry of the function  $F$  in a neighbourhood of its minimizer  $x^*$  that we recall in Section 2.

As it was shown by Attouch and Cabot in [5], if  $F$  is a strongly convex function the sequence  $(x_n)_{n \in \mathbb{N}}$  satisfies  $F(x_n) - F^* = O(n^{-\frac{2b}{3}})$  for any  $b > 0$ . In this work we give bounds depending on more general geometries than strongly convex functions, that is functions behaving like  $\|x - x^*\|^\beta$  around the minimizer for any  $\beta \geq 1$ . In particular we prove that if  $F$  is strongly convex and  $\nabla F$  is Lipschitz continuous, the decay is always better than  $F(x_n) - F^* = O(n^{-\frac{2b}{3}})$ . We also prove that the actual

decay for quadratic functions is  $F(x_n) - F^* = O(n^{-b})$ . These results rely on two geometrical conditions, one ensuring that the function is sufficiently flat around the minimizer, one ensuring it is sufficiently sharp. We recover exactly rates given in [10] for the associated ODE for any inertial parameter  $b$  and these rates are proved to be optimal in the continuous setting. Notice that since the early work of Polyak in [39], it is proven that the convergence rate of the classical gradient descent algorithm for quadratic functions is geometric. As a consequence, the decay  $O(n^{-b})$  found in the current work, for Nesterov scheme (1.5) is asymptotically slower than the one of the classical gradient method.

The current work consists in a discrete counterpart of the works [9] and [10]. Indeed, there has been a large body of literature in the 5 past years regarding the connections between dynamical systems and their discrete schemes (algorithms). In our framework it is worth mentioning the pioneering works [45] and [6] where it was shown that algorithm (1.5) is related to the following differential equation:

$$\ddot{x}(t) + \frac{b}{t}\dot{x}(t) + \nabla F(x(t)) = 0. \quad (1.6)$$

In particular, algorithm (1.5) can be seen as a time-discretization scheme of the differential equation (1.6) with a time step  $\sqrt{\gamma}$  (where  $\gamma \leq \frac{1}{L}$  corresponds to the step-size in (1.5)).

Various works have been devoted to the study of (1.6) and in particular to the convergence properties of the trajectory of solutions of (1.6), which correspond naturally to the same properties of a sequence generated by the algorithm (1.5). In particular it was shown that the convergence rate for the objective function  $F(x(t)) - F(x^*)$  is of order  $O(t^{-2})$  if  $b \geq 3$  (see [45] and [6]), while for  $b > 3$  this order is actually  $o(t^{-2})$  (see [32]) and the trajectory  $\{x(t)\}_{t>0}$  weakly converges to a minimizer of  $F$  ([6]). As for the case  $b \in (0, 3)$  this order reduces to  $O(t^{-\frac{2b}{3}})$  (see [7], [9] and [2]), and this rate is optimal [9].

Recently in [9] and [10] the authors studied the behavior of the trajectory of the solutions  $\{x(t)\}_{t>0}$  of (1.6), for a function  $F$  satisfying geometric assumptions in a neighbourhood of its minimizers (flatness and/or sharpness). The present work is thus the discrete counterpart of those papers.

The present paper is organized as follows. First we introduce in Section 2 the conditions on the local geometry of the function  $F$ . In section 3 we present the main results of this work concerning the order of convergence rate for Algorithm (1.5) and we confront them to related works. We also explain how to extend these results to a specific class of non-differential optimization problems. In section 4 we present the asymptotic analysis made for (1.5) (i.e. the proofs and the schema of the proofs of the main results). Appendix A and Appendix B contain some Lemmas (and their proofs) necessary for our analysis.

## 2 Geometry of convex functions around their minimizers

In this paragraph we present two conditions on a convex function describing its (local) geometry around the set of its minimizers. Roughly speaking, these two conditions characterize functions behaving like  $\|\cdot\|^\beta$  around its set of minimizers: one ensures that the function is sufficiently flat, while the other ensures that it is sufficiently sharp in the neighborhood of its minimizers.

**Definition 2.1.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with  $X^* = \arg \min F \neq \emptyset$ .*

1. *Let  $\beta \geq 1$ . The function  $F$  satisfies the condition  $H(\beta)$  if, for any minimizer  $x^* \in X^*$ , there exists  $\eta > 0$  such that:*

$$\forall x \in B(x^*, \eta), \quad F(x) - F(x^*) \leq \frac{1}{\beta} \langle \nabla F(x), x - x^* \rangle.$$

2. *Let  $p \geq 1$ . The function  $F$  satisfies the condition  $\mathcal{L}(p)$  if, for any minimizer  $x^* \in X^*$ , there exists a constant  $K_p > 0$  and  $\varepsilon > 0$  such that:*

$$\forall x \in B(x^*, \varepsilon), \quad K_p \|x - x^*\|^p \leq F(x) - F(x^*).$$

The hypothesis  $H(\beta)$  already used in [15, 45, 9, 10], generalizes the notion of convexity of a differentiable function in a neighborhood of its minimizers. Observe that any convex function automatically satisfies  $H(1)$ , and that any differentiable function  $F$  ensuring that  $(F - F(x^*))^{\frac{1}{\beta}}$  is convex, satisfies  $H(\beta)$  with  $\beta \geq 1$ , which is slightly more demanding than the convexity of  $F$ . To have a better insight on the local geometry of convex functions satisfying the hypothesis  $H(\beta)$ , we need the following result:

**Lemma 2.1** ([10, Lemma 2.4]). *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with  $X^* = \arg \min F \neq \emptyset$ . If  $F$  satisfies  $H(\beta)$  for some  $\beta \geq 1$ , then:*

1. *The function  $F$  satisfies  $H(\beta')$ , for all  $1 \leq \beta' \leq \beta$ .*
2. *For any minimizer  $x^* \in X^*$ , there exist  $M > 0$  and  $\eta > 0$  such that:*

$$\forall x \in B(x^*, \eta), F(x) - F(x^*) \leq M \|x - x^*\|^\beta. \quad (2.1)$$

In other words, the hypothesis  $H(\beta)$  with  $\beta \geq 1$ , can be interpreted as a flatness condition: it ensures that the function  $F$  is sufficiently flat (at least as flat as  $x \mapsto \|x\|^\beta$ ) in the neighborhood of its minimizers.

The hypothesis  $\mathcal{L}(p)$ ,  $p \geq 1$ , is a growth condition on the function  $F$  around its set of minimizers  $X^*$ . Note that, when  $X^*$  is a connected compact set, it can be replaced by a more general growth condition on  $F$  in the neighborhood of its minimizers [10], usually referred to as Hölder metric subregularity [26]:

**Lemma 2.2.** *Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}$  be a convex differentiable function satisfying the growth condition  $\mathcal{L}(p)$  for some  $p \geq 1$ . Assume that the set  $X^* = \arg \min F$  is compact. Then there exists  $K_p > 0$  and  $\varepsilon > 0$  such that, for all  $x \in \mathbb{R}^n$ :*

$$d(x, X^*) \leq \varepsilon \Rightarrow K_p d(x, X^*)^p \leq F(x) - F^*.$$

Historically, the growth condition  $\mathcal{L}(p)$  is also called  $p$ -conditioning [22] or Hölderian error bound [14], and is closely related to the Łojasiewicz inequality [30, 31], a key tool in the mathematical analysis of continuous and discrete dynamical systems [12, 13]:

**Definition 2.2.** *A differentiable function  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  is said to have the Łojasiewicz property with exponent  $\theta \in [0, 1)$  if, for any critical point  $x^*$ , there exist  $c > 0$  and  $\varepsilon > 0$  such that:*

$$\forall x \in B(x^*, \varepsilon), \|\nabla F(x)\| \geq c |F(x) - F^*|^\theta. \quad (2.2)$$

where:  $0^0 = 0$  when  $\theta = 0$  by convention.

In the convex setting, the growth condition  $\mathcal{L}(p)$ ,  $p \geq 1$ , is indeed equivalent to the Łojasiewicz inequality, with exponent  $\theta = 1 - \frac{1}{p} \in (0, 1)$  and  $c = K_p^{\frac{1}{p}}$  (see for example [12, 14]). Typical examples of functions having the Łojasiewicz property are real-analytic functions and  $C^1$  subanalytic functions or semialgebraic functions [30, 31]. Strongly convex functions satisfy a global Łojasiewicz property with exponent  $\theta = \frac{1}{2}$  [3], or equivalently a global version of the growth condition, namely:

$$\forall x \in \mathbb{R}^n, F(x) - F^* \geq \frac{\mu}{2} d(x, X^*)^2,$$

where  $\mu > 0$  denotes the parameter of strong convexity. Likewise, convex functions having a strong minimizer in the sense of [5, Section 3.3], also satisfy a global version of  $\mathcal{L}(2)$ . By extension, uniformly convex functions of order  $p \geq 2$  satisfy the global version of the hypothesis  $\mathcal{L}(p)$  (see for example [4, 14, 22] and their references).

The geometrical interpretation of the condition  $\mathcal{L}(p)$  is straightforward: it ensures that the function  $F$  is sufficiently sharp (at least as sharp as  $x \mapsto \|x - x^*\|^p$ ) in the neighborhood of its set of minimizers. Consistently, observe that any convex function satisfying  $\mathcal{L}(p)$ , satisfies  $\mathcal{L}(p')$  for all  $p' \geq p$ .

Consider now any convex differentiable function  $F$  satisfying both hypothesis  $H(\beta)$  and  $\mathcal{L}(p)$ . Combining the related inequalities, namely (2.1) and the growth condition  $\mathcal{L}(p)$ ,  $F$  has to be at least as flat as  $\|x - x^*\|^\beta$  and as sharp as  $\|x - x^*\|^p$  in the neighborhood of its minimizers.

For the simple example of the function  $F : x \in \mathbb{R} \rightarrow |x|^\gamma$  with  $\gamma > 1$ , a straightforward computation shows that  $F$  satisfies  $H(\beta)$  and  $\mathcal{L}(p)$  if and only if  $1 \leq \beta \leq \gamma \leq p$ . More generally:

**Lemma 2.3.** *If a convex differentiable function  $F$  satisfies both  $H(\beta)$  and  $\mathcal{L}(p)$ , with  $\beta, p \geq 1$ , then necessarily:  $p \geq \beta$ .*

In our framework, the objective function  $F$  is assumed to be convex and differentiable with a Lipschitz continuous gradient. For such functions, the Lipschitz continuity of the gradient provides some additional information on the local geometry of  $F$  in the neighborhood of its minimizers. Indeed, for convex functions, the Lipschitz continuity of the gradient is equivalent to a quadratic upper bound on  $F$ :

$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, F(x) - F(y) \leq \langle \nabla F(y), x - y \rangle + \frac{L}{2} \|x - y\|^2. \quad (2.3)$$

Applying (2.3) at  $y = x^*$ , we then deduce:

$$\forall x \in \mathbb{R}^N, F(x) - F^* \leq \frac{L}{2} \|x - x^*\|^2, \quad (2.4)$$

which indicates that  $F$  is at least as flat as  $\|x - x^*\|^2$  around  $X^*$ . More precisely:

**Lemma 2.4.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with a  $L$ -Lipschitz continuous gradient for some  $L > 0$ .*

1. *If  $F$  satisfies the growth condition  $\mathcal{L}(p)$ , then necessarily  $p \geq 2$ .*
2. *If  $F$  satisfies  $\mathcal{L}(2)$ , then  $F$  automatically satisfies  $H(\beta)$  with  $\beta = 1 + \frac{K_2}{2L}$  and  $K_2 \leq \frac{L}{2}$ .*

*Proof.* Assume that  $F$  satisfies the condition  $\mathcal{L}(p)$ . Combining the inequality (2.4) and the growth condition, we get: for any  $x^* \in X^*$ ,

$$K_p \|x - x^*\|^p \leq F(x) - F(x^*) \leq \frac{L}{2} \|x - x^*\|^2, \quad (2.5)$$

for all  $x$  in some neighborhood of  $x^*$ , which necessarily implies:  $p \geq 2$ . In the particular case  $p = 2$ , we also deduce that:  $2K_2 \leq L$ . The second point of Lemma 2.4 has already been shown in [10, Lemma 2.6]. □

### 3 Main results and contributions

The main results of the paper are summarized in the two theorems presented in this section. They mostly give some convergence rates for the values  $F(x_n) - F(x^*)$  where  $(x_n)_{n \in \mathbb{N}}$  is the sequence built by the Nesterov scheme (1.5) for various choices of friction parameter  $b$  and geometrical conditions  $H(\gamma)$  and  $\mathcal{L}(2)$ .

Numerous decay rates have been proposed for the Nesterov scheme [37], for gradient descent [22], or more general inertial schemes such as inertial gradient descent [5] or Heavy Ball methods [40, 27]. Some results are available for any convex functions and others assume strong convexity or condition  $\mathcal{L}(2)$ . Moreover, this inertial scheme can be seen as a discretization scheme of the specific ODE (1.6) and the given results are closely related to associated problem. We also will provide a comparison between the continuous setting and this discrete counterpart. But to be precise in all the comparisons that should be done with the state of the art we first give our results and discuss in a second time.

#### 3.1 Main results

We now present the two main Theorems of this paper. The first theorem enlightens the role of the flatness assumption  $H(\beta)$  to get better decay rates:

**Theorem 3.1.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with a  $L$ -Lipschitz continuous gradient for some  $L > 0$ . Let  $0 < \gamma \leq \frac{1}{L}$  and  $\{x_n\}_{n \in \mathbb{N}}$  be the sequence generated by Algorithm (1.5). Assume that  $F$  satisfies  $H(\beta)$  with  $\beta \geq 1$ . Then we have the following:*

1. *If  $b < 1 + \frac{2}{\beta}$ , the following convergence rates hold true asymptotically:*

$$F(x_n) - F(x^*) = O\left(n^{-\frac{2b\beta}{\beta+2}}\right) \quad \text{and} \quad \|x_n - x_{n-1}\| = O\left(n^{-\frac{b\beta}{\beta+2}}\right). \quad (3.1)$$

2. (i) If  $b \geq 1 + \frac{2}{\beta}$  then the following convergence rate holds true asymptotically:

$$F(x_n) - F(x^*) = O(n^{-2}). \quad (3.2)$$

If in addition  $b \geq 2$ , then  $\{x_n\}_{n \geq 1}$  is bounded and:

$$\|x_n - x_{n-1}\| = O(n^{-1}). \quad (3.3)$$

(ii) If  $b > 1 + \frac{2}{\beta}$  then:

$$\sum_{n=0}^{+\infty} n(F(x_n) - F(x^*)) < +\infty \quad \text{and} \quad \sum_{n=0}^{+\infty} n\|x_n - x_{n-1}\|^2 < +\infty. \quad (3.4)$$

In addition the sequence  $\{x_n\}_{n \in \mathbb{N}}$  converges to a minimizer  $x^*$ .

(iii) If  $b \geq 1 + \frac{2}{\beta}$  and if  $F$  satisfies  $\mathcal{L}(2)$  and admits a unique minimizer, then the following convergence rates hold true asymptotically:

$$F(x_n) - F(x^*) = O\left(n^{-\frac{2b\beta}{\beta+2}}\right) \quad \text{and} \quad \|x_n - x_{n-1}\| = O\left(n^{-\frac{b\beta}{\beta+2}}\right). \quad (3.5)$$

**Corollary 3.1.** In view of point 2(ii) of Theorem 3.1, in the case  $b > 1 + \frac{2}{\beta}$ , for  $n \in \mathbb{N}$ , we actually have:

$$F(x_n) - F(x^*) = o(n^{-2}) \quad \text{and} \quad \|x_n - x_{n-1}\| = o(n^{-1}), \quad (3.6)$$

as also the sequence  $x_n$  converges to a minimizer  $x^*$ .

This corollary is an extension of Theorem 1 in [8]. The proof of this corollary is a direct consequence of the summability (3.4) and of Lemma A.2 in Appendix A (these two results are key elements of the convergence proof in [17]) as detailed hereafter.

We define  $U_n = F(x_n) - F(x^*) + \frac{\|x_n - x_{n-1}\|^2}{2\gamma}$ . From (3.4) we deduce that

$$\sum_{n \geq 1} nU_n < +\infty. \quad (3.7)$$

It follows that for any  $\varepsilon > 0$ , it exists a rank  $n_0$  such that for any  $n \geq n_0$ ,

$$\sum_{k=\lfloor \frac{n}{2} \rfloor}^n kU_k < \varepsilon. \quad (3.8)$$

Hence, the sequence  $\tilde{U}_n = \min_{k \in [\lfloor \frac{n}{2} \rfloor, n]} U_k$  satisfies

$$\tilde{U}_n \times \sum_{k=\lfloor \frac{n}{2} \rfloor}^n k \leq \sum_{k=\lfloor \frac{n}{2} \rfloor}^n kU_k < \varepsilon. \quad (3.9)$$

and thus  $\tilde{U}_n = o\left(\frac{1}{n^2}\right)$ . Moreover  $(U_n)_{n \geq 1}$  is non-increasing (see Lemma A.2 in Appendix A) and thus  $U_n = \tilde{U}_n$  which concludes the proof of (3.6).

The proof of the convergence of the sequence  $x_n$  to a minimizer  $x^*$  is also based on the estimates (3.4) and it is identical to the one made in [17] (see Theorem 3) and is omitted here.

Let us make some observations: first, the point 2(iii) of Theorem 3.1 only applies for  $\beta \leq 2$ , since there exists no function satisfying both  $H(\beta)$  with  $\beta > 2$ , and  $\mathcal{L}(2)$  (see Lemma 2.3). Note also that the uniqueness of the minimizer is needed to apply the point 2(iii) of Theorem 3.1, but it is still an open problem to know if this uniqueness assumption is necessary. In a context such as the Least Square problem, the hypothesis on the uniqueness of  $x^*$  can be omitted since the whole trajectory  $(x_n)_{n \in \mathbb{N}}$  belongs to an affine space where the solution of the minimization problem is unique.

**Remark 1 (The Least Square problem).** Let us consider the Least square inverse problem. Given  $y \in \mathbb{R}^M$  and a positive-definite bounded linear operator (matrix)  $A : \mathbb{R}^N \rightarrow \mathbb{R}^M$ , we consider the

function  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  such that  $F(x) = \frac{1}{2}\|Ax - y\|^2$  for all  $x \in \mathbb{R}^N$  and the minimization problem (1.1), i.e.:

$$\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - y\|^2. \quad (\text{LS})$$

For the problem (LS), the algorithm (1.5) reads:

$$\begin{aligned} y_n &= x_n + \alpha_n(x_n - x_{n-1}) \\ x_{n+1} &= y_n - \gamma A^*(Ay_n - y) \end{aligned} \quad (3.10)$$

with  $x_0 = x_1 \in \mathbb{R}^N$ . We then deduce that for all  $n \in \mathbb{N}$ ,  $x_n \in \{x_0\} + \text{Im}A^*$ .

Since this problem has a unique solution on the space  $\{x_0\} + \text{Im}A^*$ , the second point of Theorem 3.1 is applicable.

The second main Theorem of this work deals with flat geometries, i.e. with functions satisfying  $H(\beta)$  and  $\mathcal{L}(p)$  with  $p \geq \beta > 2$ , in the neighborhood of their minimizer:

**Theorem 3.2.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with a  $L$ -Lipschitz continuous gradient for some  $L > 0$ . Let  $0 < \gamma \leq \frac{1}{L}$  and  $\{x_n\}_{n \in \mathbb{N}}$  be the sequence generated by Algorithm (1.5).*

*Assume that  $F$  satisfies  $H(\beta)$  and  $\mathcal{L}(p)$  with  $p \geq \beta > 2$ , and that  $F$  has a unique minimizer. If  $b \geq \frac{\beta+2}{\beta-2}$ , then the following estimate hold true:*

$$F(x_n) - F(x^*) = O\left(n^{-\frac{2p}{p-2}}\right).$$

In Table 1 we give an overall summary of the results in Theorems 3.1 and 3.2, concerning the convergence properties of Nesterov's scheme (1.5), depending on the over-relaxation parameter  $b > 0$  and the local geometry of the function  $F$  (i.e. parameters  $\beta$  and  $p$ ).

$b$ (inertia)	$H(\beta)$ and $\mathcal{L}(p)$	rates $F(x_n) - F^*$	convergence
$b \leq 1 + \frac{2}{\beta}$	$H(\beta)$	$O\left(n^{-\frac{2\beta b}{\beta+2}}\right)$	
$b > 1 + \frac{2}{\beta}$	$H(\beta)$	$o(n^{-2})$	$x_n \rightarrow x^*$
$b \geq 1 + \frac{2}{\beta}$	$H(\beta)$ & $\mathcal{L}(2)$ uniqueness of the minimizer	$O\left(n^{-\frac{2\beta b}{\beta+2}}\right)$	$x_n \rightarrow x^*$
$b \geq \frac{\beta+2}{\beta-2}$	$H(\beta)$ & $\mathcal{L}(p)$ , $p \geq \beta > 2$ uniqueness of the minimizer	$O\left(n^{-\frac{2p}{p-2}}\right)$	$x_n \rightarrow x^*$

Table 1: Summary of the convergence properties of Nesterov's scheme (1.5).

Theorem 3.1 enlightens the role of the flatness hypothesis  $H(\beta)$ . As in the continuous setting [10], we will see that this flatness hypothesis may ensure a better decay rate, avoiding too large oscillations around the minimizer.

Let us first recall some state of the art results about the Nesterov family of inertial schemes (1.5). It is known since [36] in 1983 that for  $b = 3$  the Nesterov scheme (1.5) ensures that  $F(x_n) - F(x^*) = O\left(\frac{1}{n^2}\right)$ , if  $F$  is convex. Actually, the proof of Nesterov was also available for  $b \geq 3$  but the value  $b = 3$  ensures the lowest constant hidden in the big O. That is why this choice was used for the generalization to non smooth functions FISTA [11] or for restarting methods [38]. In [17], Chambolle and Dossal reminded that any  $b \geq 3$  may provide the same decay, and they showed that if  $b > 3$ , we can moreover ensure the weak convergence of iterates in a Hilbert space and the convergence of  $\sum_{n \in \mathbb{N}} n(F(x_n) - F(x^*))$ . In [8], Attouch and Peypouquet deduced from this summability that  $F(x_n) - F(x^*) = o\left(\frac{1}{n^2}\right)$  when  $b > 3$ . More recently, following similar results from the continuous setting, Attouch et al. [7] and Apidopoulos et al. [1] proved that  $F(x_n) - F(x^*) = O\left(n^{-\frac{2b}{3}}\right)$  when  $b \leq 3$ . Points 1, 2(i) and 2(ii) from Theorem 3.1 highlight that if  $F$  satisfies  $H(\beta)$  for  $\beta > 1$ , the decay rate is actually better than  $O\left(n^{-\frac{2b}{3}}\right)$  for all  $b \leq 3$  and that the limit value for parameter  $b$  ensuring the convergence of iterates and the convergence of  $\sum_{n \in \mathbb{N}} n(F(x_n) - F(x^*))$  is actually  $1 + \frac{2}{\beta} < 3$ . As a consequence for such functions, the sequence of iterates converges for  $b = 3$ .



For classical gradient descent, the condition  $\mathcal{L}(2)$  is the key to get the optimal decay rate, since the function  $F(x) = |x|^p$  satisfies  $\mathcal{L}(p)$  by definition and achieves the best decay rate for functions satisfying  $\mathcal{L}(p)$  [33]. But, considering inertial algorithms, we will see that the sharpness is not always the key assumption, and that to get better rates for Nesterov scheme (1.5), one need to control both the sharpness and the flatness of the function.

Consider first the class of strongly convex functions. In that case, Su et al. proved in [45] that for  $b \leq 9$ , if  $F$  is strongly convex then:  $F(x_n) - F(x^*) = O\left(n^{-\frac{2b}{3}}\right)$ . This result was extended to any  $b > 0$  by Attouch et al. [8] for functions having a strong minimizer. The strong convexity or the strong minimizer assumption ensure that  $H(1)$  and a global version of  $L(2)$  are in force. So, applying the point 2(iii) of Theorem 3.1, we recover the rate  $O\left(n^{-\frac{2b}{3}}\right)$  stated in [45, 8]. Now, if  $F$  satisfies a flatness condition  $H(\beta)$  for some  $\beta > 1$  and only  $L(2)$ , then the rate given by the point 2(iii) of Theorem 3.1 is then strictly better than  $O\left(n^{-\frac{2b}{3}}\right)$ .

If  $F$  is a quadratic function in the neighborhood of its minimizer, then  $F$  naturally satisfies  $L(2)$  and  $H(\beta)$  for any  $\beta \in [1, 2]$ . Consequently applying the point 2(iii) of Theorem 3.1 with  $\beta = 2$ , we prove that the decay rate is  $O(n^{-b})$  for quadratic functions, which is better than  $O\left(n^{-\frac{2b}{3}}\right)$  that can be obtained using only strong convexity. Since this rate is proved to be optimal for the continuous setting [10], we conjecture that it is optimal in the discrete setting as well. Note that, in a continuous setting, under the sole assumption  $\mathcal{L}(2)$ , we cannot get a better rate than  $\frac{2b}{3}$ . A flatness hypothesis is needed to get the power  $b$ .

Finally, if  $F$  is a convex differentiable function with a Lipschitz continuous gradient, and if  $F$  satisfies  $L(2)$ , then  $F$  automatically satisfies  $H(\beta)$  for some  $\beta \in ]1, 2]$  (see Lemma 2.4). In that case Point 2(iii) of Theorem 3.1 applies, showing that for such functions the decay is always faster than  $O\left(n^{-\frac{2b}{3}}\right)$ .

Theorem 3.2 deals with the case when  $F$  satisfies  $\mathcal{L}(p)$  with  $p > 2$  but not necessarily  $\mathcal{L}(2)$ . To the best of our knowledge, there are no references with convergence results for the Nesterov's algorithms family (1.5), under these kinds of assumptions. This theorem ensures that if  $b$  is large enough, the decay of  $F(x_n) - F(x^*)$  is faster than the rate  $o(n^{-2})$  that can be obtained only with a convexity assumption. For some similar results, concerning a restarting version of Algorithm (1.5), for functions satisfying  $H(\beta)$ -kind conditions (in particular relation (2.1) of Lemma 2.1) and  $\mathcal{L}(p)$ , with  $\beta < p$ , we address the reader to [35] and [42].

**Comparison with gradient descent and other inertial algorithms** Many inertial schemes of gradient descent have been proposed such as the Inertial Gradient Descent (1.5)

$$x_{n+1} = x_n + \alpha_n(x_n - x_{n-1}) - \gamma \nabla F(x_n + \alpha_n(x_n - x_{n-1})), \quad (3.11)$$

or the Heavy Ball Algorithm

$$x_{n+1} = x_n + \alpha_n(x_n - x_{n-1}) - \gamma \nabla F(x_n). \quad (3.12)$$

The case  $\alpha_n = 0$  of (3.11) corresponds to the classical gradient descent, while  $\alpha_n = \frac{n}{n+b}$  is the case studied in this article. If  $b = 3$  we recover the original choice of Nesterov, but other choices of over-relaxation sequence  $\alpha_n$  have been studied, see for example [5].

Under the hypothesis  $\mathcal{L}(2)$ , the gradient descent or other general descent schemes ensure a geometrical decay of  $F(x_n)$  to  $F^*$  (see for example [14] and [22]). This decay is also geometrical for the Heavy Ball method and the inertial gradient descent with fixed inertial parameter for strongly convex functions [37] (see also [38], [34]). More precisely Nesterov in 1983 [36] proposed to choose a fixed sequence  $\alpha_n = \frac{\sqrt{L} - \sqrt{\alpha}}{\sqrt{L} + \sqrt{\alpha}}$  where  $L$  is the Lipschitz constant of the gradient of  $F$  and  $\alpha$  is the parameter of strong convexity of  $F$ , to optimize this geometrical decay. Estimating numerically  $Q = \frac{L}{\alpha}$  can be a very challenging task. There exists a vast body of literature on adapting restarting versions of inertial algorithms in order to estimate the conditional number  $Q$  (to cite but a few of these works, we address the reader to [38], [19], [20], [29], [34] [16] and [42]). We can notice that for inertial methods the uniqueness of the minimizer is needed, which is not the case for gradient descent. Notice also that in the first points of Theorem 3.1, we do not assume the hypothesis  $\mathcal{L}(2)$  or any strong convexity. Theorem 3.1 ensures that if  $b > 1 + \frac{2}{\beta}$  the decay of  $F(x_n) - F(x^*)$  is faster than  $O\left(\frac{1}{n}\right)$  which seems to be the best bound we can achieve with such hypotheses for gradient descent or other inertial methods such as the heavy ball algorithm.

In the recent work [25] the authors propose a different application of alternated inertia to (1.5) (i.e. applying the inertial term every two iterations), in a non smooth setting. Surprisingly this turns Algorithm (1.5) with alternated inertia, into a descent scheme and it permits to have the same convergence properties as the Forward-Backward algorithm under the hypothesis  $\mathcal{L}(p)$  with  $p \geq 1$ .

In a recent work, Attouch et al. [5] proved that the decay is faster than polynomial for a large class of inertial parameters  $\alpha_n$  such that  $1 - \alpha_n \sim \frac{1}{n^r}$  with  $r \in [0, 1)$ .

While Theorem 3.1 improves the bound for Nesterov scheme (1.5) on  $F(x_n) - F(x^*)$  that can be obtained with the sole assumption  $\mathcal{L}(2)$ , by adding the condition  $H(\beta)$ , these decays are still polynomials and they are thus worse than the ones of the classical gradient decent, the inertial gradient descent with fixed parameter, or the Heavy Ball algorithm. Nevertheless it is possible that for a given precision, the Nesterov scheme (1.5) can potentially have better performance than the gradient descent, in finite time, by tuning properly the parameter  $b$ . Of course in order to do such a comparison, one must have access to the explicit dependence on the hidden constants in the "big Oh's" of the over-relaxation parameter  $b > 0$  or/and the geometric parameters  $\beta$  and  $p$ . Unfortunately our analysis is highly based on asymptotic equivalences, which does not permit to have an explicit formulation of these constants and this question is left for future study.

For this issue we are also addressing the reader to Figure 1, where we test Nesterov's scheme (1.5) for three simple minimization problems (Least squares, Quadratic and Least squares with Tikhonov regularization), that enter in the framework of Theorem 1. More precisely we compare the Gradient Descent algorithm (black) with the Nesterov scheme (1.5), with three different choices for the over-relaxation parameter  $b$  ( $b = 4$  blue,  $b = 20$  red and  $b = 80$  green).

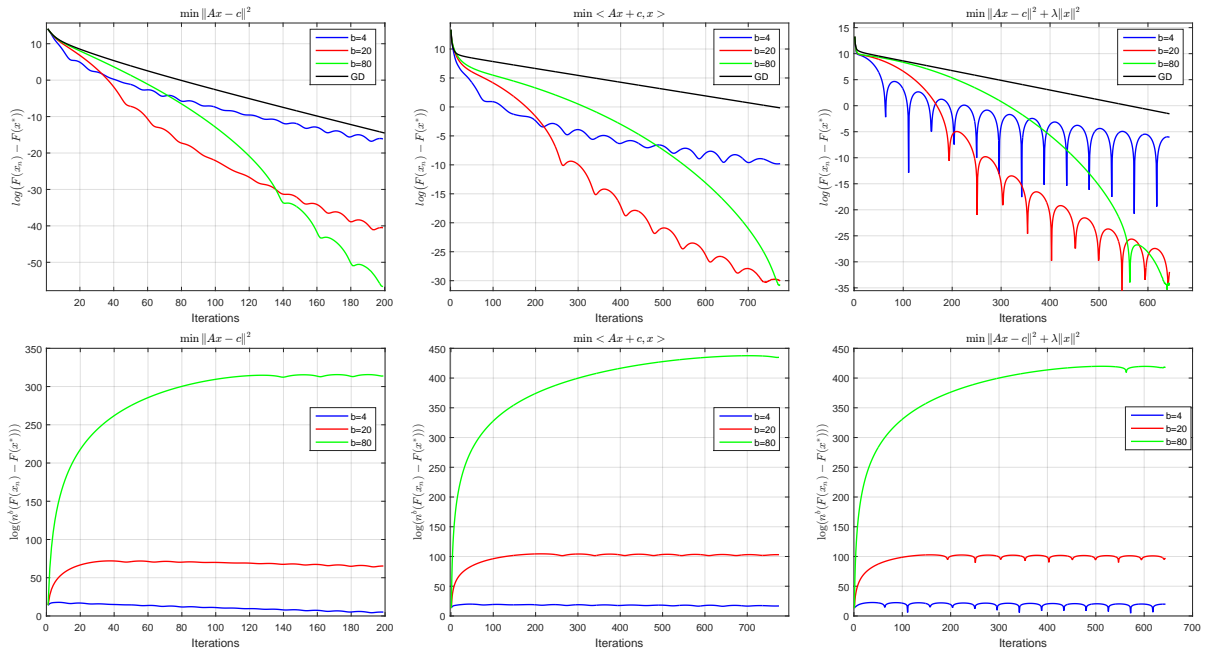


Figure 1: Values of the error  $\log(F(x_n) - F(x^*))$  (first row) and of the energies  $\log n^b(F(x_n) - F(x^*))$  (second row), as a function of the iterations  $n$ , for the three minimization problems (Least squares, Quadratic and Tikhonov regularization). Here we compare the Nesterov scheme (1.5) with 3 different choices of  $b$  ( $b = 4, 10, 80$ ) and the Gradient-Descent. Notice that smaller values of  $b$  seem to lead to a better performance in the beginning, while the larger values accelerate the convergence rate at a later stage. From the second column we can also notice that the re-scaled errors  $\log n^b(F(x_n) - F(x^*))$  seem to be bounded, which suggests that the rate  $O(n^{-b})$  found in Theorem 1 is tight.

From Figure 1, one can notice that for all the three problems, the over-relaxation parameter  $b$  plays a crucial role for the convergence rate of the objective function. Indeed larger values of  $b$ , seem to lead to faster convergence rates "at a late stage", as Theorem 3.1 asserts. Nevertheless, it is also worth mentioning that taking  $b$  very large, results to "slower" convergence behavior at the beginning. This last remark suggests that the over-relaxation parameter  $b$ , has also a serious impact in the hidden constants in the "big Oh's" of the different estimates in Theorem 1. This entails that for given a precision (or a

stopping time), choosing  $b > 0$  arbitrarily large may not be always an optimal strategy.

Finally, Theorem 3.2 ensures that under the assumption  $\mathcal{L}(p)$  with  $p > 2$ , if  $b$  is large enough, the decay of  $F(x_n) - F(x^*)$  is  $O(n^{-\frac{2p}{p-2}})$  which is better than  $o(n^{-2})$  and better than what can be obtained with the sole hypothesis  $\mathcal{L}(p)$  for gradient descent schemes. Indeed under the assumption  $\mathcal{L}(p)$ , Bolte et al. [14] (see also [21] and [22] and their references) proved that the gradient descent ensures a decay which is  $O(n^{-\frac{2}{p-2}})$  and this decay is optimal under this assumption as proven in [33]. Unfortunately in our work the optimality of the order  $O(n^{-\frac{2p}{p-2}})$  is not proven. Nevertheless for functions such as  $\|x\|^p$ ,  $p > 2$ , this order seems to be tight as Figure 2 propounds.

In particular, in Figure 2, we are considering the simple example of minimizing the function  $F : \mathbb{R}^4 \rightarrow \mathbb{R}_+$ , such that  $F(x) = \|x\|^p$ , with  $p > 2$ . In this setting we compare the i-GD Algorithm with four different choices for the over-relaxation parameter  $b$  (blue, green, red and magenta) for different values of  $p$  ( $p = 3$ ,  $p = 4$ , and  $p = 8$ ).

In this case we intentionally choose a parameter  $b = \frac{p+1}{p-2} < \frac{p+2}{p-2}$  (blue line) that violates the assumptions of Theorem 2. This seems to cause an overshoot on the minimum of the corresponding trajectory. On the contrary the trajectories that respect the hypotheses of Theorem 3.2 with  $b \geq \frac{p+2}{p-2}$  (red, green and magenta lines), seem to produce a decreasing behavior, without bump effects. From Figure 2, we can notice that the re-scaled errors  $n^{\frac{2p}{p-2}} (F(x_n) - F^*)$ , seem to be bounded. This is not the case for the blue line which corresponds to a choice of  $b < \frac{p+2}{p-2}$  which does not enter in the framework of Theorem 2. This fact suggests that the order of convergence rate found in Theorem 3.2 for the Nesterov scheme (1.5) is optimal for this kind of functions, under the assumption  $b \geq \frac{p+2}{p-2}$ . In addition comparing all the three trajectories that are bounded, the best bound is obtained by the limiting case of the trajectory which corresponds to  $b = \frac{p+2}{p-2}$ . This suggests that the limiting value  $b = \frac{p+2}{p-2}$  may minimize the hidden constants in the "big Oh" of Theorem 3.2.

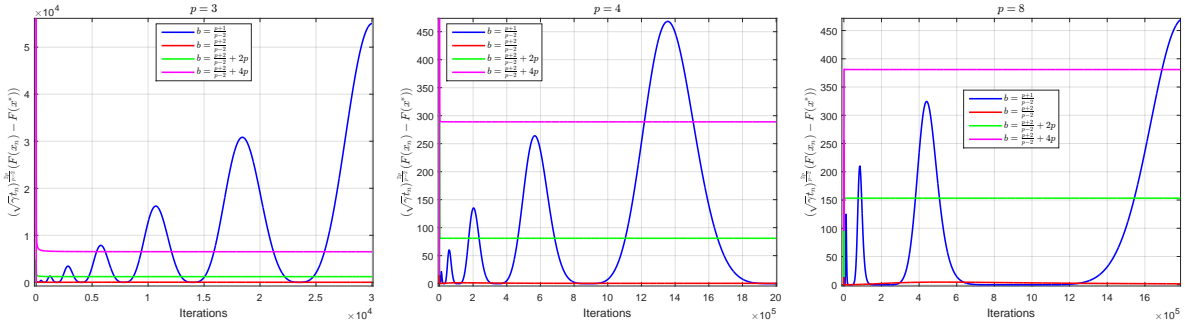


Figure 2: Values of the re-scaled error  $F(x_n) - F(x^*)$ , by  $(\sqrt{\gamma}(n+b-1))^{\frac{2p}{p-2}}$  as a function of the iterations  $n$ . Here we test the Nesterov scheme (1.5) for four different values of  $b$  depending on  $p$  ( $b = \frac{p+1}{p-2}$  blue,  $b = \frac{p+2}{p-2}$  red,  $b = \frac{p+2}{p-2} + 2p$  green and  $b = \frac{p+2}{p-2} + 4p$  magenta). Each column correspond to the minimization of  $\|x\|^p$ , with  $p = 3$ ,  $p = 4$  and  $p = 8$  respectively. Here we can remark that all the three choices of parameter  $b$  (red, green and magenta) that enter the framework of Theorem 2, are bounded. On the contrary the blue one seems to explode periodically along the iterations. Notice also that among the three bounded lines, the one that has a better upper bound (red), corresponds to the smallest possible value of  $b$  that enters the framework of Theorem 2 (i.e.  $b = \frac{p+2}{p-2}$ )

**Comparison with the continuous setting:** Both theorems of this paper can be seen as extension of previous results [9, 10] on the solution of the ODE associated to the Nesterov Scheme (1.5). In [45], Su et al. were the first to propose results for both continuous and discrete setting. After this article most recent results on Inertial Gradient Descent have been developed first in a continuous setting, that is studying solutions of ODEs such that

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \nabla F(x(t)) = 0. \quad (3.13)$$

The main reason is the higher complexity of the discrete setting where some discretization terms must be bounded. The choice of the discretization may be crucial. One can notice that the Heavy Ball Method and the Inertial Gradient Descent with constant parameter  $\alpha_n = \alpha$  are associated to the same ODE

(3.13) with a constant friction term  $\alpha(t) = \alpha$ . Nevertheless, both algorithms do not share exactly the same properties and the same optimal tuning for parameters see [27] and [44]. The hypothesis of both theorems are similar to the ones in the continuous setting, with the Lipschitz hypothesis on the gradient of  $F$ . We also need uniqueness of the minimizer for Theorem 3.2. We think that this hypothesis could be removed but the technical price to pay was too high and we prefer to assume this uniqueness to keep a readable proof.

### 3.2 Extension to the non-differential setting

Many interesting problems in optimization are structured by minimizing a non-differentiable convex function  $F$ . Typically, a classical type of these problems can be formulated by minimizing a function  $F$ , which takes the form  $F = f + g$ , where  $f$  and  $g$  are both convex, with  $f$  differentiable with  $L$ -Lipschitz gradient and  $g$  lower semi-continuous (possibly non-differentiable) and such that  $\arg \min F \neq \emptyset$ .

In this setting (with  $x_0 = x_1 \in \mathbb{R}^N$  and  $0 < \gamma \leq \frac{1}{L}$ ), the algorithm (1.5) takes the following form :

$$\begin{aligned} y_n &= x_n + \frac{n}{n+b}(x_n - x_{n-1}) \\ x_{n+1} &= T_\gamma(y_n) := \text{Prox}_{\gamma g}(y_n - \gamma \nabla f(y_n)) := \left( \text{Id} + \gamma \partial g \right)^{-1} (y_n - \gamma \nabla f(y_n)). \end{aligned} \quad (3.14)$$

Unfortunately the main core of this paper does not handle with such a framework which is let for future study. Nevertheless in the case of convex functions  $f$  and  $g$ , the hypothesis  $H(1)$  is automatically satisfied. In view of the proof of Lemma A.1 (see also Lemma 1, in [17]), our analysis is still applicable. Therefore, from Theorem 3.1 we can recover some of the results concerning the algorithm (3.14), already found in previous works ( see for example [1], [5], [6], [7],[8], [17]). In particular we have the following Corollary.

**Corollary 3.2.** *Let  $F = f + g$ , where  $f$  and  $g$  are convex, lower semi-continuous functions, with  $f$  differentiable with  $L$ -Lipschitz gradient, such that  $\arg \min F \neq \emptyset$ . Let also  $0 < \gamma \leq \frac{1}{L}$  and  $\{x_n\}_{n \in \mathbb{N}}$  be the sequence generated by Algorithm (3.14).*

1. *If  $b < 3$ , the following convergence rates hold true asymptotically:*

$$F(x_n) - F(x^*) = O\left(n^{-\frac{2b}{3}}\right) \quad \text{and} \quad \|x_n - x_{n-1}\| = O\left(n^{-\frac{b}{3}}\right). \quad (3.15)$$

2. (i) *If  $b \geq 3$  then the following convergence rate holds true asymptotically:*

$$F(x_n) - F(x^*) = O(n^{-2}) \quad \text{and} \quad \|x_n - x_{n-1}\| = O(n^{-1}). \quad (3.16)$$

- (ii) *If  $b > 3$  then:*

$$\sum_{n=0}^{+\infty} n(F(x_n) - F(x^*)) < +\infty \quad \text{and} \quad \sum_{n=0}^{+\infty} n\|x_n - x_{n-1}\|^2 < +\infty. \quad (3.17)$$

*In fact :  $F(x_n) - F(x^*) = o(n^{-2})$  and  $\|x_n - x_{n-1}\| = o(n^{-1})$  and the sequence  $\{x_n\}_{n \in \mathbb{N}}$  converges to a minimizer  $x^*$ .*

- (iii) *If  $b \geq 3$  and if  $F$  satisfies  $\mathcal{L}(2)$  and admits a unique minimizer, then the following convergence rate holds true asymptotically:*

$$F(x_n) - F(x^*) = O\left(n^{-\frac{2b}{3}}\right) \quad \text{and} \quad \|x_n - x_{n-1}\| = O\left(n^{-\frac{b}{3}}\right). \quad (3.18)$$

As we shall remark, Corollary 3.2 can be applied to a class of problems of a particular interest such as the (generalized) LASSO problem. In that case, the minimizing function  $F$  is convex piecewise polynomial of degree 2, hence (see Corollary 9 in [14]), for all  $r \geq F^*$ , the function  $F$  satisfies  $\mathcal{L}(2)$  on every sub-level set  $\{F \leq r\} = \{x \in \mathbb{R}^N : F(x) \leq r\}$ . This inducts that the point 2.(iii) of Corollary 3.2 is applicable, under the supplementary condition of the uniqueness of a minimizer of  $F$ . More precisely we have the following Corollary.

**Corollary 3.3** (Generalized LASSO). *Let  $F(x) = \frac{1}{2}\|Ax - y\|^2 + \lambda\|Tx\|_1$ , where  $A : \mathbb{R}^N \rightarrow \mathbb{R}^M$  and  $T : \mathbb{R}^N \rightarrow \mathbb{R}^M$  are some linear operators,  $y \in \mathbb{R}^M$  and  $\lambda > 0$ . Suppose also that  $\arg \min F = \{x^*\}$ . Let  $\{x_n\}_{n \in \mathbb{N}}$ , be the sequence generated by algorithm (3.14). Then for all  $b > 0$ , the following convergence rate holds true asymptotically :*

$$F(x_n) - F(x^*) = O\left(n^{-\frac{2b}{3}}\right).$$

*Remark 2.* More generally (see Corollary 4.3 in [18] or [22]), if  $F(x) = f(Ax) + g(Tx)$ , with some linear operators  $A \in \mathbb{R}^{N \times M}$  and  $T \in \mathbb{R}^{N \times M}$  and  $f, g$  are convex piecewise-linear quadratic functions (see Definition 10.20 in [41]), then  $F$  satisfies  $\mathcal{L}(2)$ . Hence by assuming the uniqueness of the minimizer, the point 2.(iii) of Corollary 3.2 is still applicable in this more general setting.

Last but not least, we shall mention that for the minimization problem of  $F = f + g$ , where  $f, g$  are both convex with  $f$  differentiable with  $L$ -Lipschitz gradient and  $g$  lower semi-continuous, by making additional hypotheses, such as non-degeneracy condition on  $F$  and restricted injectivity on  $f$  (for more details see the work [28]), the uniqueness of the minimizer  $x^*$  is reassured. If in addition  $g$  is partially smooth at  $x^*$  relative to a neighborhood of  $x^*$  (see Definition 5 in [28]), then  $F$  satisfies  $\mathcal{L}(2)$  locally in  $x^*$ . Thus, in that case we can recover the order of  $O\left(n^{-\frac{2b}{3}}\right)$  of the point 2.(iii) of Corollary 3.2.

## 4 Asymptotic analysis

In this section we give the proofs of Theorems 3.1 and 3.2. Before passing to the complete proofs, we provide the necessary tools, as also a basic sketch in order to have a better insight.

### 4.1 Schema of proofs

The basic tools that we use are a Lyapunov-type analysis and the asymptotic equivalences. The choice of the Lyapunov energy-sequence, as also the asymptotic analysis are highly-inspired by the work made in the continuous-time counterpart for a solution of (1.6) in the work [10].

In this context, Lyapunov techniques consist in finding a suitable positive energy-sequence  $E_n$  of the form:

$$E_n = \varphi_n(F(x_n) - F(x^*)) + R_n \quad (4.1)$$

with  $\varphi_n$  and  $R_n$  some positive sequences. Then by showing that  $E_n$  is non-increasing, it follows directly that the convergence rate for the objective  $F(x_n) - F(x^*)$  is of order of  $O(\varphi_n^{-1})$ .

In this work, we set  $\varphi_n \sim n^2$  and the term  $R_n$  is not necessarily non-negative. In fact the exact construction of  $E_n$  (see (4.12)), depends on the geometric properties of  $F$  and on the order of the convergence rate of the objective function, i.e. the value of  $\delta$  such that:

$$F(x_n) - F(x^*) = O(n^{-\delta}) \quad (4.2)$$

as stated in Theorems 3.1 and 3.2.

In order to get the estimation (4.2), a two-step procedure is used:

1. First of all, since  $\varphi_n \sim n^2$ , we show the control over the growth or the decay of  $E_n$  of the following form (see for example relation (4.18) for Theorem 3.1 and (4.36) for Theorem 3.2):

$$E_n \leq K n^{-\delta+2} \quad (4.3)$$

2. Since  $R_n$  is not necessarily non-negative we cannot deduce (4.2) directly from (4.3). For this issue, we infer the geometric properties of  $F$  (in particular hypothesis  $\mathcal{L}(p)$ ) in order to deduce (4.2) from (4.3).

To get the appropriate control (4.3) on  $E_n$ , we follow a classical strategy for bounding functions using a differential inequality, which is motivated by the continuous-time setting (see [10]).

In fact for a differentiable function  $\mathcal{E} : [0 + \infty) \rightarrow \mathbb{R}$ , a constant  $c \in \mathbb{R}$  and positive function  $r : [0 + \infty) \rightarrow \mathbb{R}_+$ , such that the following relation holds true:

$$\mathcal{E}'(t) \leq \frac{c\mathcal{E}(t)}{t} + r(t) \quad (4.4)$$

with  $t^{-c}r(t) \in L^1[0, +\infty)$ , one can deduce that the function  $H(t) = t^{-c}\mathcal{E}(t)$  is bounded from above. Thus it exists a constant  $K \in \mathbb{R}$  such that  $\mathcal{E}(t) \leq Kt^c$ .

More precisely, in the framework of the current work (discrete setting), we provide an energy-sequence  $\{E_n\}_{n \geq 1}$ , such that the following relation holds true asymptotically:

$$E_{n+1} - E_n \leq \frac{c}{n}E_n + r_n \quad (4.5)$$

with  $c = -\delta + 2$  and a suitable sequence  $\{r_n\}_{n \geq 1}$  that involves the geometric properties of  $F$ . In particular, in the context of Theorem 3.1, we have that  $r_n = \frac{a}{n^2}E_n$  (see relation (4.17) of Lemma 4.2), while  $r_n = \frac{a_1}{n^2}E_n + \frac{a_2}{n^2}\|x_{n-1} - x^*\|^2$  (see relation (4.28) of Lemma 4.3), for some suitable positive constants  $a$ ,  $a_1$  and  $a_2$ . This allows to deduce the existence of a constant  $K$  such that for  $n \in \mathbb{N}$  large enough, we have:

$$E_n \leq Kn^c \quad (4.6)$$

Finally, in order to deduce (4.2) from (4.3) we use different strategies depending on the hypotheses on the geometry of  $F$  and on the over-relaxation parameter  $b$ .

1. For the first point of Theorem 3.1, the sequence  $\{R_n\}$  in (4.1) is positive and (4.2) holds directly.
2. For the second point of Theorem 3.1 and for the bound of Theorem 3.2, we have  $R_n = R'_n + \xi\|x^* - x_n\|^2$ , where  $R'_n$  is non negative and  $\xi$  non positive. Thus, from (4.3), for  $n$  large enough, it follows that:

$$\varphi_n(F(x_n) - F(x^*)) - |\xi|\|x_n - x^*\|^2 \leq Kn^{-\delta+2} \quad (4.7)$$

and we conclude, using the growth condition  $L(p)$  to bound  $\|x^* - x_n\|^2$  and get inequalities such that

$$\varphi_n(F(x_n) - F(x^*)) + A_1(F(x_n) - F(x^*))^{\frac{2}{p}} \leq Kn^{-\delta+2} \quad (4.8)$$

which, by recalling that  $\varphi_n = n^2$  and using an appropriate strategy when  $p > 2$  (see Lemma A.5), leads to (4.2).

The value of  $\delta$  and the use of conditions  $L(p)$  are different in the two theorems, which leads to different results, but the strategies in both cases are similar.

## 4.2 The Lyapunov energy

We now give the proper definition of the energy-sequence  $\{E_n\}_{n \geq 1}$ , in order to proceed to the analysis described before. Let us introduce some notations that will be useful in the coming analysis:

$$w_n = F(x_n) - F(x^*), \quad \delta_n = \|x_n - x_{n-1}\|^2 \quad \text{and} \quad h_n = \|x_n - x^*\|^2. \quad (4.9)$$

For some  $\lambda > 0$  and  $\xi \in \mathbb{R}$ , we also define:

$$v_n = \|\lambda(x_{n-1} - x^*) + t_n(x_n - x_{n-1})\|^2, \quad n \geq 1, \quad (4.10)$$

with

$$t_n = n + b - 1 \quad \text{and} \quad \alpha_n = \frac{n}{t_{n+1}} \quad (4.11)$$

and:

$$\begin{aligned} E_n &= (t_n^2 + \lambda\beta t_n)w_n + \frac{1}{2\gamma}\|\lambda(x_{n-1} - x^*) + t_n(x_n - x_{n-1})\|^2 + \frac{\lambda t_n}{2\gamma}\|x_n - x_{n-1}\|^2 + \frac{\xi}{2\gamma}\|x_{n-1} - x^*\|^2 \\ &= (t_n^2 + \lambda\beta t_n)w_n + \frac{1}{2\gamma}v_n + \frac{\lambda t_n}{2\gamma}\delta_n + \frac{\xi}{2\gamma}h_{n-1} \end{aligned} \quad (4.12)$$

Observe that the energy can also be expressed as:

$$E_n = (t_n^2 + \lambda\beta t_n)w_n + \frac{1}{2\gamma} \left( t_n^2 \delta_n + \lambda t_n (h_n - h_{n-1}) + (\lambda^2 + \xi) h_{n-1} \right) \quad (4.13)$$

*Remark 3.* By definition of  $E_n$  and using the convex inequality:

$$\|u\|^2 \leq 2\|u+v\|^2 + 2\|v\|^2, \quad \forall u, v \in \mathbb{R}^N \quad (4.14)$$

with  $u = t_n(x_n - x_{n-1})$  and  $v = \lambda(x_{n-1} - x^*)$ , we find:

$$\begin{aligned} 2\gamma E_n &\geq 2\gamma(t_n^2 + \lambda\beta t_n)w_n + \left(\frac{t_n^2}{2} + \lambda t_n\right)\|x_n - x_{n-1}\|^2 + (\xi - \lambda^2)\|x_{n-1} - x^*\|^2 \\ &\geq 2\gamma t_n^2 w_n + \frac{t_n^2}{2}\delta_n + (\xi - \lambda^2)h_{n-1} \end{aligned} \quad (4.15)$$

In what follows we frequently make use of the inequality (4.15).

We also give the following basic Lemma which gives some bound estimates for the energy  $E_n$  and will be useful for both of the proofs of Theorem 3.1 and Theorem 3.2.

**Lemma 4.1.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with a  $L$ -Lipschitz continuous gradient for some  $L > 0$ . Let  $0 < \gamma \leq \frac{1}{L}$  and  $\{x_n\}_{n \in \mathbb{N}}$  be the sequence generated by Algorithm (1.5). Assume that  $F$  satisfies  $H(\beta)$  with  $\beta \geq 1$  and  $x^* \in \arg \min F$ . Then for all  $\lambda \geq 0$  and  $\xi = \lambda(\lambda + 1 - b)$  in the definition of  $E_n$ , the following recursive formula holds for all  $n \geq 1$ :*

$$2\gamma(E_{n+1} - E_n) \leq 2\gamma \frac{c(\lambda)}{t_n} E_n + 2\gamma \left( A_1(\lambda)t_{n+1} - 2\lambda\beta(\lambda + 1 - b) \right) w_n + A_2(\lambda)\|x_n - x_{n-1}\|^2 + \frac{A_3(\lambda)}{t_n}\|x_{n-1} - x^*\|^2 \quad (4.16)$$

where:

$$\begin{aligned} c(\lambda) &= 2(\lambda + 1 - b), \quad A_1(\lambda) = 2b - (\beta + 2)\lambda, \quad A_2(\lambda) = (2\lambda + 1 - b)(1 - b) \\ A_3(\lambda) &= -2\lambda(\lambda + 1 - b)(2\lambda + 1 - b) \end{aligned}$$

### 4.3 Proofs

Here we present the complete proofs of Theorems 3.1 and 3.2 as also the different Lemmas that form the guideline described before. For ease of reading, all the proofs of these Lemmas are postponed in the Appendix B.

#### 4.3.1 Proof of Theorem 3.1

Before proving Theorem 3.1, we present a Lemma which first gives the control estimates over the local variation of the energy  $\{E_n\}_{n \geq 1}$  depending on the parameters  $\beta$  and  $b$ , and then a control estimate of  $E_n$ .

**Lemma 4.2.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with a  $L$ -Lipschitz continuous gradient for some  $L > 0$ . Let  $0 < \gamma \leq \frac{1}{L}$  and  $\{x_n\}_{n \in \mathbb{N}}$  be the sequence generated by Algorithm (1.5). Assume that  $F$  satisfies  $H(\beta)$  with  $\beta \geq 1$ , and that one of the following hypotheses is in force:*

- i.  $b < 1 + \frac{2}{\beta}$
- ii.  $b \geq 1 + \frac{2}{\beta}$  and  $F$  admits a unique minimizer  $x^*$  and satisfies  $\mathcal{L}(2)$ .

Taking  $\lambda = \frac{2b}{\beta+2}$  and  $\xi = \lambda(\lambda + 1 - b) = \frac{2b\beta}{(\beta+2)^2} \left(1 + \frac{2}{\beta} - b\right)$  in the definition of the energy  $E_n$ , we have:

1. There exists some  $n_0 \in \mathbb{N}$ , such that for all  $n \geq n_0$ , the following recursive formula holds true:

$$E_{n+1} - E_n \leq \left( \frac{a}{(n+b-1)^2} + \frac{c}{(n+b-1)} \right) E_n \quad (4.17)$$

for some constant  $a \geq 0$  and  $c = 2 - \frac{2b\beta}{\beta+2}$ .

2. The following estimate holds true asymptotically:

$$E_n = O\left(n^{2 - \frac{2b\beta}{\beta+2}}\right) \quad (4.18)$$

We are now ready to give the complete proof of Theorem 3.1

*Proof of Theorem 3.1.* We start this demonstration by proving the points 1 and 2(iii) of Theorem 3.1. For that, we choose:

$$\lambda = \frac{2b}{\beta + 2} > 0, \quad \xi = \lambda(\lambda + 1 - b) = \frac{2b\beta}{(\beta + 2)^2} \left(1 + \frac{2}{\beta} - b\right),$$

in the definition (4.12) of the energy  $E_n$ . Using Lemma 4.2, there exist  $n_0 \in \mathbb{N}$  and a positive constant  $C$  such that, for all  $n \geq n_0$ , we have:

$$E_n \leq Ct_n^{2 - \frac{2b\beta}{\beta + 2}}. \quad (4.19)$$

In order to deduce the expected convergence rates on  $w_n = F(x_n) - F(x^*)$ , we use different strategies depending on the sign of the parameter  $\xi$ . Firstly we consider the case  $b < 1 + \frac{2}{\beta}$ , i.e.:  $\xi > 0$ . In that case, the energy  $E_n$  is a sum of non-negative terms, hence:

$$E_n = (t_n^2 + \lambda\beta t_n)w_n + \frac{1}{2\gamma}(v_n + \lambda t_n \delta_n + \xi h_{n-1}) \geq t_n^2 w_n.$$

Combining the very last inequality with (4.19) and noting that  $t_n \sim n$  asymptotically, we get the appropriate estimate:  $w_n = O\left(n^{-\frac{2b\beta}{\beta + 2}}\right)$  for all  $n \geq n_0$ , as asserted by the first point of Theorem 3.1.

In addition, since  $\xi > 0$ , from the definition of the energy  $E_n$  (4.12) and (4.18) we find :

$$h_{n-1} \leq \frac{2\gamma}{\xi} E_n \leq K t_n^{2 - \frac{2b\beta}{\beta + 2}} \quad (4.20)$$

asymptotically, for some suitable positive constant  $K$ .

Finally, using the inequality (4.15), we have:

$$\frac{t_n^2}{2} \delta_n \leq 2\gamma E_n - 2\gamma t_n^2 w_n + (\lambda^2 - \xi)h_{n-1} \leq 2\gamma E_n + |\lambda^2 - \xi|h_{n-1}. \quad (4.21)$$

Injecting estimations (4.19) and (4.20) into (4.21) leads to:  $\delta_n = O\left(t_n^{-\frac{2b\beta}{\beta + 2}}\right)$  as expected.

Consider now the case  $b \geq 1 + \frac{2}{\beta}$  i.e.  $\xi \leq 0$ . In that case, the energy  $E_n$  is not a sum of non negative terms anymore:

$$2\gamma E_n = 2\gamma(t_n^2 + \lambda\beta t_n)w_n + v_n + \lambda t_n \delta_n - |\xi||x_{n-1} - x^*|^2,$$

and an additional growth condition  $\mathcal{L}(2)$  will be needed to bound  $\|x_{n-1} - x^*\|^2$ . First, applying the inequality (4.14), on the one hand to  $u = t_n(x_n - x_{n-1})$  and  $v = \lambda(x_{n-1} - x^*)$  and on the other hand to  $u = x_{n-1} - x^*$  and  $v = x^* - x_n$ , we have for all  $n \in \mathbb{N}$ :

$$v_n \geq \frac{t_n^2}{2} \delta_n - \lambda^2 \|x_{n-1} - x^*\|^2, \quad \|x_{n-1} - x^*\|^2 \leq 2\delta_n + 2h_n.$$

Using these two inequalities successively, we deduce:

$$\begin{aligned} 2\gamma E_n &\geq 2\gamma(t_n^2 + \lambda\beta t_n)w_n + \left(\frac{1}{2} + \frac{\lambda}{t_n}\right)t_n^2 \delta_n + (\xi - \lambda^2)\|x_{n-1} - x^*\|^2 \\ &\geq 2\gamma(t_n^2 + \lambda\beta t_n)w_n + \left(\frac{1}{2} + \frac{\lambda}{t_n} - \frac{2|\xi - \lambda^2|}{t_n^2}\right)t_n^2 \delta_n - 2|\xi - \lambda^2|\|x_n - x^*\|^2 \\ &\geq 2\gamma(t_n^2 + \lambda\beta t_n)w_n + \frac{t_n^2}{4} \delta_n - 2|\xi - \lambda^2|\|x_n - x^*\|^2 \end{aligned} \quad (4.22)$$

since the coefficient of  $t_n^2 \delta_n$  converges to  $\frac{1}{2}$ , and thus is greater than e.g.  $\frac{1}{4}$  for  $n$  large enough. Assuming in addition that  $F$  satisfies the growth condition  $\mathcal{L}(2)$  and admits a unique minimizer, we then obtain for  $n$  large enough:

$$E_n \geq (t_n^2 + \lambda\beta t_n - 2|\xi - \lambda^2|K_2^{-1})w_n + \frac{t_n^2}{4} \delta_n = \left(1 + \frac{\lambda\beta}{t_n} - 2\frac{|\xi - \lambda^2|K_2^{-1}}{t_n^2}\right)t_n^2 w_n + \frac{t_n^2}{4} \delta_n$$

Hence there exists  $n_0 \in \mathbb{N}$  such that for all  $n \geq n_0$ , we have:  $E_n \geq \frac{t_n^2}{2} w_n + \frac{t_n^2}{4} \delta_n$ . Using finally the estimate (4.19) on the energy  $E_n$  allows us to conclude the proof of the point 2(iii) of Theorem 3.1.



Consider again the case when  $b \geq 1 + \frac{2}{\beta}$ . In order to prove the points 2(i) and 2(ii) of Theorem 3.1 with the only assumption that  $F$  satisfies the condition  $H(\beta)$ , we choose a different value for the parameter  $\lambda$  in the definition of the energy  $E_n$ . Let us set:

$$\lambda = b - 1 > 0, \quad \xi = \lambda(\lambda + 1 - b) = 0.$$

In that case, the energy is again a sum of non negative terms and we then have:

$$E_n = (t_n^2 + (b-1)\beta t_n)w_n + \frac{1}{2\gamma}(v_n + (b-1)t_n\delta_n) \geq t_n^2 w_n. \quad (4.23)$$

To obtain the expected convergence rate on  $w_n$  as expressed in the point 2(i) of Theorem 3.1, it is sufficient to prove that the energy  $E_n$  is bounded. For that purpose, we apply Lemma 4.1 with  $\lambda = b - 1$ . Keeping in mind that  $b \geq 1 + \frac{2}{\beta}$ , we then have:

$$\begin{aligned} \forall n \geq 1, E_{n+1} - E_n &\leq (\beta(1 + \frac{2}{\beta} - b)t_n + 1)w_n - \frac{1}{2\gamma}(b-1)^2\delta_n \\ &\leq w_n - \frac{1}{2\gamma}(b-1)^2\delta_n \leq w_n \end{aligned} \quad (4.24)$$

Injecting (4.23) into (4.24), we then obtain for all  $n \geq 1$ ,  $E_{n+1} \leq (1 + \frac{1}{t_n^2})E_n$ , which implies by a recurrence argument that:

$$\forall n \geq 1, E_{n+1} \leq E_1 \prod_{i=1}^n (1 + \frac{1}{t_i^2})$$

By inferring Lemma A.4 with  $c = 0$  and  $a = 1$  we deduce that the sequence  $(E_n)_{n \geq 1}$  is bounded. By (4.23), the sequence  $(t_n^2 w_n)_{n \geq 1}$  is also bounded. Hence:  $w_n = O(t_n^{-2}) = O(n^{-2})$  asymptotically.

Assume in addition that:  $b \geq 2$ . According to the definition (4.13) of the energy  $E_n$ , we have:

$$\begin{aligned} \forall n \geq 1, 2\gamma E_n &= 2\gamma(t_n^2 + (b-1)\beta t_n)w_n + t_n^2\delta_n + (b-1)t_n(h_n - h_{n-1}) + (b-1)^2 h_{n-1} \\ &\geq (b-1)t_n(h_n - h_{n-1}) + (b-1)^2 h_{n-1} \\ &\geq (b-1)(t_n h_n - (t_n - b + 1)h_{n-1}) \end{aligned}$$

Observe now that, since  $b \geq 2$ , we have:  $t_n - b + 1 \leq t_n - 1 = t_{n-1}$ , hence:

$$\forall n \geq 1, 2\gamma E_n \geq (b-1)(t_n h_n - t_{n-1} h_{n-1}).$$

Using the fact that  $(E_n)_{n \geq 1}$  is bounded, there exist a constant  $C > 0$  and  $n_0 \in \mathbb{N}$  such that:

$$\forall n \geq n_0, t_n h_n - t_{n-1} h_{n-1} \leq C. \quad (4.25)$$

By summing (4.25) from  $n_0$  to  $N$ , we obtain that for all  $N \geq n_0$ ,  $t_N h_N \leq t_{n_0} h_{n_0} + CN \leq t_{n_0} h_{n_0} + Ct_N$ . The sequence  $(h_n)_n$  is so bounded, which implies that  $(x_n)_n$  is also bounded. Moreover using (4.15):

$$\forall n \geq 1, \frac{t_n^2}{2}\delta_n \leq 2\gamma E_n + \lambda^2 h_{n-1},$$

and the boundedness of  $h_n$  and  $E_n$ , we obtain the boundedness of the sequence  $(t_n^2 \delta_n)_n$ , i.e.  $\delta_n = O(n^{-2})$  asymptotically, which concludes the proof of point 2(i) of Theorem 3.1.

Finally, suppose that  $b > 1 + \frac{2}{\beta}$ . Let:  $\eta = b - (1 + \frac{2}{\beta}) > 0$ . As previously done in (4.24), we have:

$$E_{n+1} - E_n \leq -(\beta\eta t_n - 1)w_n - \frac{1}{2\gamma}(b-1)^2\delta_n \leq -(\beta\eta t_n - 1)w_n \quad (4.26)$$

Moreover using the fact that there exists  $n_0 \in \mathbb{N}$  such that for all  $n \geq n_0$ , we have:  $\beta\frac{\eta}{2}t_n \leq \beta\eta t_n - 1$  and summing (4.26) over  $n \in \{n_0, \dots, N\}$ , for all  $N > n_0$ , we find:

$$\beta\frac{\eta}{2} \sum_{n=n_0}^N t_n w_n \leq E_{n_0} < +\infty \quad (4.27)$$

Lastly, the proof of the summability of the  $(n\delta_n)_n$  is exactly the same as in [17, Corollary 2]. In few words, applying Lemma A.1, with  $\gamma \leq \frac{1}{L}$ ,  $y = x_n + \alpha_n(x_n - x_{n-1})$  and  $x = x_n$ , or equivalently using the Lipschitz continuity of the gradient of  $F$ , we have:

$$\forall n \geq 1, \delta_{n+1} - \alpha_n^2 \delta_n \leq 2\gamma(w_n - w_{n+1}),$$

where:  $\alpha_n = \frac{n}{n+b}$ . Summing the last inequality from  $n = 1$  to  $N$ , we obtain:

$$(b-1) \sum_{n=2}^N (2t_n - b + 1) \delta_n \leq 2\gamma \sum_{n=2}^N (2t_n + 1)w_n + t_2^2 w_1.$$

Observe now that for all  $n \geq 1$ , we have:  $2t_n - b + 1 = 2n + b - 1 \geq 2n$ . Thus the summability of  $n\delta_n$  follows from the summability of  $t_n w_n$ , which concludes the proof of point 2(ii) of Theorem 3.1.  $\square$

### 4.3.2 Proof of Theorem 3.2

Let us now consider the case of a function  $F$  which satisfies  $H(\beta)$  and  $\mathcal{L}(p)$  with  $\beta > 2$  and  $p > 2$ . In order to give the complete proof of Theorem 3.2, we make use of the following Lemma.

**Lemma 4.3.** *Let  $F : \mathbb{R}^N \rightarrow \mathbb{R}$  be a convex differentiable function with a  $L$ -Lipschitz continuous gradient for some  $L > 0$ . Let  $0 < \gamma \leq \frac{1}{L}$  and  $\{x_n\}_{n \in \mathbb{N}}$  be the sequence generated by Algorithm (1.5).*

*Assume that  $F$  satisfies  $H(\beta)$  and  $\mathcal{L}(p)$  with  $p \geq \beta > 2$ . Let  $\lambda = \frac{2}{\beta-2}$  and  $\xi = \lambda(\lambda + 1 - b) = \frac{2}{\beta-2} \left( \frac{\beta}{\beta-2} - b \right)$  in the definition of the energy  $E_n$ . If  $b \geq \frac{\beta+2}{\beta-2}$ , then there exist  $n_0 \in \mathbb{N}$  and some constant  $C > 0$  such that the following recursive formula holds true for all  $n \geq n_0$ :*

$$2\gamma(E_{n+1} - E_n) \leq 2\gamma \left( -\frac{4}{\beta-2} + \frac{C}{t_n} \right) \frac{E_n}{t_n} + \frac{C}{t_n^2} h_{n-1}. \quad (4.28)$$

We are now ready to give the full proof of Theorem 3.2.

*Proof of Theorem 3.2.* We split the proof into three parts. In the first part we present the analysis in order to obtain a control over the decay of the energy  $E_n$ . In the second part we deduce some estimates for the sequence  $w_n$ , by making use of an *a-priori* estimate for  $h_n$ . In the last part we infer a bootstrap argument to improve the estimates for the sequence  $w_n$  and to get those stated in Theorem 3.2.

*Part 1:* In this part, we show that choosing  $\lambda = \frac{2}{\beta-2}$  and  $\xi = \lambda(\lambda + 1 - b) = \frac{2}{\beta-2} \left( \frac{\beta}{\beta-2} - b \right)$  in the definition of the energy  $E_n$ , the control over the decay of the energy  $E_n$  is given by:

$$E_n = O(n^{-m}), \quad \text{with: } m = \begin{cases} \frac{4}{p} & \text{if } \beta = \frac{6p+8}{p+2} \\ \min\left(\frac{4}{\beta-2}, 1 + \frac{4}{p}\right) & \text{otherwise} \end{cases}.$$

In this proof we will frequently use the notation  $d = \frac{4}{\beta-2}$ . By Lemma 4.3, there exist  $n_0 \in \mathbb{N}$  and a positive constant  $C$  such that

$$\forall n \geq n_0, E_{n+1} - E_n \leq -\frac{d}{t_n} E_n + \frac{C}{t_n^2} E_n + \frac{C}{2\gamma t_n^2} h_{n-1}. \quad (4.29)$$

Denoting by  $H_n = \frac{C}{2\gamma t_n^2} h_{n-1}$  and  $z_n = 1 - \frac{d}{t_n} + \frac{C}{t_n^2}$ , the previous inequality (4.29) can be rewritten as:

$$E_{n+1} \leq z_n E_n + H_n \quad (4.30)$$

which, by applying Lemma A.3, implies:

$$\forall n \geq n_0, E_{n+1} \leq \prod_{i=n_0}^n z_i \left( E_{n_0} + \sum_{i=n_0}^n \frac{H_i}{\prod_{m=n_0}^i z_m} \right). \quad (4.31)$$

By relation (A.18) of Lemma A.4, we deduce the existence of two positive constants  $C_1$  and  $C_2$  such that for all  $n \geq n_0$  it holds:

$$C_1 t_n^{-d} \leq \prod_{i=n_0}^n z_i = \prod_{i=n_0}^n \left(1 - \frac{d}{t_i} + \frac{C}{t_i^2}\right) \leq C_2 t_n^{-d}. \quad (4.32)$$

Hence by (4.31) and the definition of  $H_i = \frac{C}{2\gamma t_i^2} h_{i-1}$ , we find:

$$\begin{aligned} E_{n+1} &\leq C_2 t_n^{-d} \left( E_{n_0} + C_1^{-1} \sum_{i=n_0}^n t_i^d H_i \right) \\ &\leq C_2 t_n^{-d} \left( E_{n_0} + C_3 \sum_{i=n_0}^n t_i^{d-2} h_{i-1} \right) \end{aligned} \quad (4.33)$$

for some suitable positive constant  $C_3$ . So to obtain an estimate on the energy, we first need an estimate on  $h_n$ . Assuming that  $F$  satisfies the growth condition  $\mathcal{L}(p)$  with  $p > 2$ , and admits a unique minimizer  $x^*$ , we have:

$$h_{i-1} = \|x_{i-1} - x^*\|^2 \leq K_p^{-1} (F(x_{i-1}) - F(x^*))^{\frac{2}{p}} = K_p^{-1} w_{i-1}^{\frac{2}{p}}. \quad (4.34)$$

By injecting the last inequality (4.34) into (4.33), for  $n \geq n_0$  we find:

$$\begin{aligned} E_{n+1} &\leq C_2 t_n^{-d} \left( E_{n_0} + C_3 K_p^{-1} \sum_{i=n_0}^n t_i^{d-2} w_{i-1}^{\frac{2}{p}} \right) \\ &\leq C_2 t_n^{-d} \left( E_{n_0} + C_3 K_p^{-1} \sum_{i=n_0}^n t_i^{d-2-\frac{4}{p}} (t_i^2 w_{i-1})^{\frac{2}{p}} \right) \end{aligned} \quad (4.35)$$

Moreover, since  $F$  satisfies  $H(\beta)$  with  $\beta > 2$ , we can apply the results stated in Theorem 3.1. Here:  $b \geq \frac{\beta+2}{\beta-2} > 1 + \frac{2}{\beta}$ . Hence from relation (3.4) of Theorem 3.1, the sequence  $(t_i^2 w_{i-1})_i$  is bounded. Therefore, from the previous inequality (4.35), by using the series-integral comparison test, we find that for all  $n \geq n_0$ , it holds:

$$E_{n+1} \leq C_2 t_n^{-d} \left( E_{n_0} + C_4 t_n^{\max\{d-1-\frac{4}{p}, 0\}} \right), \quad \text{if } d \neq 1 + \frac{4}{p}$$

or :

$$E_{n+1} \leq C_2 t_n^{-d} \left( E_{n_0} + C_4 \log t_n \right) \leq C_2 t_n^{-d} \left( E_{n_0} + C_4 t_n \right), \quad \text{if } d = 1 + \frac{4}{p}$$

for some suitable positive constant  $C_4$  (at each case). Therefore there exists a suitable positive constant  $C > 0$ , such that :

$$E_{n+1} \leq C t_n^{-m} \quad (4.36)$$

for  $m = \begin{cases} \frac{4}{p} & \text{if } d = 1 + \frac{4}{p} \\ \min\{d, 1 + \frac{4}{p}\} & \text{otherwise} \end{cases}$ .

*Part 2:* Once obtained the control (4.36) over the decay of  $E_n$ , we now want to deduce the convergence rates on  $w_n = F(x_n) - F(x^*)$  expected in Theorem 3.2.

Firstly, observe that when  $\beta > 2$  and  $b \geq \frac{\beta+2}{\beta-2}$ , we have

$$\lambda = \frac{2}{\beta-2} > 0 \quad \text{and} \quad \xi = \lambda(\lambda+1-b) = \frac{2}{\beta-2} \left( \frac{\beta}{\beta-2} - b \right) < 0.$$

The energy  $E_n$  is not a sum of non negative terms, so that, as in Theorem 3.1, we will so need some growth condition to bound  $\|x_{n-1} - x^*\|^2$ , or more precisely  $\|x_n - x^*\|^2$  in what follows. First remark that using (4.15), we get:

$$\begin{aligned} 2\gamma E_n &= 2\gamma(t_n^2 + \lambda\beta t_n)w_n + v_n + \lambda t_n \delta_n - |\xi| \|x_{n-1} - x^*\|^2 \\ &\geq 2\gamma t_n^2 w_n + \frac{t_n^2}{2} \delta_n - \lambda(b-1) \|x_{n-1} - x^*\|^2. \end{aligned} \quad (4.37)$$

By using the inequality

$$\|x_{n-1} - x^*\|^2 \leq 2\|x_n - x^*\|^2 + 2\|x_n - x_{n-1}\|^2$$

in (4.37) we find:

$$2\gamma t_n^2 w_n + \left(\frac{t_n^2}{2} - 2\lambda(b-1)\right)\delta_n \leq 2\gamma E_n + 2\lambda(b-1)h_n. \quad (4.38)$$

Hence, there exists  $n_0 \in \mathbb{N}$  such that for  $n \geq n_0$ :

$$\begin{aligned} t_n^2 w_n &\leq 2\gamma E_n + 2\lambda(b-1)\|x_n - x^*\|^2 \\ &\leq C t_n^{-m} + 2\lambda(b-1)\|x_n - x^*\|^2 \end{aligned}$$

using the control estimate (4.36) on the energy  $E_n$  for some suitable positive constant  $C$ . Using the growth condition  $\mathcal{L}(p)$  with  $p > 2$  combined with the uniqueness of the minimizer, gives:

$$t_n^2 w_n \leq C t_n^{-m} + \frac{2\lambda(b-1)}{K_p} w_n^{\frac{2}{p}} \quad (4.39)$$

Deducing now the convergence rates on  $w_n$  is quite technical: multiplying (4.39) by  $t_n^m$  and setting  $g_n = t_n^{m+2} w_n$ , we find:

$$g_n \leq C + \frac{2\lambda(b-1)}{K} t_n^{\frac{mp-2(m+2)}{p}} g_n^{\frac{2}{p}} \quad (4.40)$$

By applying Lemma A.5 with  $z_n = \frac{2\lambda(b-1)}{K} t_n^{\frac{mp-2(m+2)}{p}}$  and  $\alpha = \frac{2}{p} \in (0, 1)$ , we obtain:

$$\begin{aligned} g_n &\leq 2 \max \left\{ C, \left( 2^{\frac{2}{p}} \frac{2\lambda(b-1)}{K} t_n^{\frac{mp-2(m+2)}{p}} \right)^{\frac{p}{p-2}} \right\} = 2 \max \left\{ C, C' t_n^{m - \frac{4}{p-2}} \right\} \\ &= O(t_n^M) \end{aligned} \quad (4.41)$$

where

$$M = \max \left\{ 0, m - \frac{4}{p-2} \right\} = \begin{cases} m - \frac{4}{p-2} & \text{if } p \geq \max(\beta, 4) \text{ and } d \neq 1 + \frac{4}{p} \\ 0 & \text{otherwise.} \end{cases}$$

Substituting then  $g_n = t_n^{m+2} w_n$ , we finally have:

$$w_n = O(t_n^{M-m-2}). \quad (4.42)$$

At this point we consider the different disjoint cases for the parameters  $(\beta, p)$  in order to precise the estimate (4.42). Recalling that  $\beta \leq p$ ,  $m = \begin{cases} \frac{4}{p} & \text{if } d = 1 + \frac{4}{p} \\ \min\{d, 1 + \frac{4}{p}\} & \text{otherwise} \end{cases}$  and  $M = \max\{0, m - \frac{4}{p-2}\}$  we have the following cases: Let us first suppose that  $d = \frac{4}{\beta-2} \neq 1 + \frac{4}{p}$

- If  $m = d = \frac{4}{\beta-2}$  then (since  $\beta \leq p$ ), we have necessarily that  $M = m - \frac{4}{p-2}$ . In that case by (4.42), we find:  $w_n = O\left(t_n^{-\frac{2p}{p-2}}\right)$

- If  $m = 1 + \frac{4}{p}$  then:

- If  $M = m - \frac{4}{p-2}$ , then from (4.42) we find:  $w_n = O\left(t_n^{-\frac{2p}{p-2}}\right)$

- If  $M = 0$ , from (4.42) we find:  $w_n = O\left(t_n^{-(3+\frac{4}{p})}\right)$

Lastly, if  $d = 1 + \frac{4}{p}$ , then  $m = \frac{4}{p} < \frac{4}{p-2}$ , thus  $M = 0$  and from (4.42), we find :  $w_n = O\left(t_n^{-(2+\frac{4}{p})}\right)$

The previous cases can be regrouped into two regimes  $\mathcal{B}_1$  and  $\mathcal{B}_2$ , for the parameters  $(\beta, p) \in \{(x, y) \in \mathbb{R}^2 : 2 < x \leq y\}$  with:

$$\mathcal{B}_1 : \{p \geq 4\} \cap \{d \neq 1 + \frac{4}{p}\}$$

$$\mathcal{B}_2 : \left( \{p \leq 4\} \cap \{d > 1 + \frac{4}{p}\} \right) \cup \{d = 1 + \frac{4}{p}\}$$

such that:

- If  $(\beta, p) \in \mathcal{B}_1$  then from (4.42) we obtain:

$$w_n = O\left(t_n^{-\frac{2p}{p-2}}\right) \quad (4.43)$$

- If  $(\beta, p) \in \mathcal{B}_2$  then from (4.42) we obtain:

$$w_n = O(t_n^{-\mu_1}) \quad \text{with } \mu_1 = \begin{cases} 2 + \frac{4}{p} & \text{if } d = 1 + \frac{4}{p} \\ 3 + \frac{4}{p} & \text{otherwise} \end{cases} \quad (4.44)$$

In the case when  $(\beta, p) \in \mathcal{B}_1$ , we can conclude directly the proof of Theorem 3.2.

Let us now treat the case of  $(\beta, p) \in \mathcal{B}_2$ . In this case the estimate found in (4.44) is sub-optimal, in comparison with the one stated in Theorem 3.2. This point is also strongly accented by the corresponding results for the continuous-time version (see Theorem 4.3 in [10]). This is due to the use of the *a-priori* estimate for  $(t_n^2 w_{n-1})_n$  used in (4.35), in our analysis.

Nevertheless, we show that this estimate can be "improved" by inferring a bootstrap argument for a suitable amount of times. More precisely the idea is to use (4.44) as an *a-priori* estimate, by re-injecting it in (4.34). This idea is presented in the third part.

*Part 3:* First we define the sequences  $\{\mu_l\}_{l \in \mathbb{N}}$ ,  $\{m_l\}_{l \in \mathbb{N}}$  and  $\{M_l\}_{l \in \mathbb{N}}$ , with  $\mu_0 = 2$ , such that for all  $l \geq 1$  it holds:

$$\mu_l = \begin{cases} 2 + \frac{2}{p}\mu_{l-1} & \text{if } \mu_{l-1} = \frac{p(d-1)}{2} \\ 3 + \frac{2}{p}\mu_{l-1} & \text{otherwise} \end{cases} \quad (4.45)$$

and for all  $l \in \mathbb{N}$ :

$$m_l = \begin{cases} \frac{2}{p}\mu_l & \text{if } d = 1 + \frac{2}{p}\mu_{l-1} \\ \min\{d, 1 + \frac{2}{p}\mu_l\} & \text{otherwise} \end{cases} \quad \text{and} \quad M_l = \max\left\{0, m_l - \frac{4}{p-2}\right\} \quad (4.46)$$

(note that  $m_0 = m$  and  $M_0 = M$ ).

For all  $l \in \mathbb{N}$ , we also define the following family of conditions  $\mathcal{B}_2(l)$ :

$$\mathcal{B}_2(l) : \quad M_l = 0 \quad \text{and} \quad m_l = \begin{cases} \frac{2}{p}\mu_l & \text{if } d = 1 + \frac{2}{p}\mu_l \\ 1 + \frac{2}{p}\mu_l & \text{otherwise} \end{cases}. \quad (4.47)$$

Since  $\mathcal{B}_2(0)$  is in force, by relation (4.44) we have that:

$$w_n = O(t_n^{-\mu_1}) \quad (4.48)$$

Hence, by using the hypothesis  $\mathcal{L}(p)$  and the uniqueness of the minimizer, we find that:

$$h_n \leq K_2^{-1} w_n^{\frac{2}{p}} = O\left(t_n^{-\frac{2}{p}\mu_1}\right) \quad (4.49)$$

By following the same procedure as before in Part 1 and injecting the inequality (4.49) into (4.33), and using the series-integral test, we find:

$$E_{n+1} = O(t_n^{-m_1}) \quad (4.50)$$

By proceeding exactly in the same way as before in Part 2, one can deduce that:

$$w_n = O(t_n^{M_1 - m_1 - 2}) \quad (4.51)$$

If we suppose that  $\mathcal{B}_2(1)$  does not hold true (i.e.  $M_1 = m_1 - \frac{4}{p-2}$ ), then the result of Theorem 3.2 follows directly from relation (4.51). If in the contrary  $\mathcal{B}_2(1)$  is in force, then from (4.51), it follows that:

$$w_n = O(t_n^{-\mu_2}) \quad (4.52)$$

In fact, in the same way as before, by a recurrence argument, we find that for all  $l \geq 0$  it holds:

$$w_n = O(t_n^{M_l - m_l - 2}) \quad (4.53)$$

and if  $\mathcal{B}_2(l)$  holds true, then :  $w_n = O\left(t_n^{-\mu_{l+1}}\right)$ , otherwise  $w_n = O\left(t_n^{-\frac{2p}{p-2}}\right)$ .

Let us prove by contradiction that  $\mathcal{B}_2(l)$  cannot hold true for all  $l \in \mathbb{N}$ . For that we suppose that the condition  $\mathcal{B}_2(l)$ , holds true for all  $l \in \mathbb{N}$ . In that case for all  $l \in \mathbb{N}$ , since  $M_l = 0$ , we have  $m_l < \frac{4}{p-2}$  which is equivalent to:

$$\left\{\mu_l \leq \frac{(6-p)p}{2(p-2)} \text{ and } \mu_l \neq \frac{(d-1)p}{2}\right\} \text{ or } \left\{\mu_l \leq \frac{2p}{p-2} \text{ and } \mu_l = \frac{(d-1)p}{2}\right\} \quad (4.54)$$

Notice in that case that  $\mu_l$  is always an increasing sequence, converging to its supremum:  $\frac{3p}{p-2}$ . In fact one can assure that there is at most one  $l_1 \in \mathbb{N}$ , such that  $\mu_{l_1} = \frac{(d-1)p}{2} \leq \frac{2p}{p-2}$ , thus  $\mu_{l+1} = 3 + \frac{2}{p}\mu_l$  for all  $l \geq l_1$ . This entails that after a certain rank  $l^* \in \mathbb{N}$ , for all  $l \geq l^*$ , we have  $\mu_l > \frac{2p}{p-2}$ , which together with (4.54), leads to a contradiction.

Thus, we deduce the existence of an  $l^* \in \mathbb{N}$ , such that  $\mathcal{B}_2(l^*)$  does not hold true. Therefore (since the condition  $\mathcal{B}_2(l^*)$  does not hold true), by (4.53) we deduce that:

$$w_n = O\left(t_n^{-\frac{2p}{p-2}}\right) \quad (4.55)$$

which concludes the proof of Theorem 3.2.  $\square$

## A General Lemmas

In this section we give some auxiliary lemmas that we use in our analysis.

First we give a basic descent-type lemma for the function  $F$ , concerning the operator  $T_\gamma(x) := x - \gamma F(x)$  (see also Lemma 2.2 in [11] or Lemma 1, in [17]).

**Lemma A.1.** *Let  $\gamma > 0$  and  $F$  satisfying  $H(\beta)$  with  $\beta \geq 1$ . For every  $(x, y) \in \mathbb{R}^N$  we have that:*

$$2\gamma(F(T_\gamma(y)) - F(x)) \leq \|y - x\|^2 - \|T_\gamma(y) - x\|^2 + (\gamma L - 1)\|T_\gamma(y) - y\|^2 \quad (A.1)$$

In addition for all  $y \in \mathbb{R}^N$  and  $x^* \in X^*$  it holds:

$$2\gamma(F(T_\gamma(y)) - F(x^*)) \leq \frac{1}{\beta} \left( \|y - x^*\|^2 - \|T_\gamma(y) - x^*\|^2 \right) + \left( \gamma L + \frac{1}{\beta} - 2 \right) \|T_\gamma(y) - y\|^2 \quad (A.2)$$

*Proof.* The first point is already settled in previous works (see for example Lemma 2.3 in [11] or Lemma 1, in [17] for the proximal setting). Nevertheless we recall the complete proof of it.

Using the fact that  $\nabla F$  is  $L$ -Lipschitz, for all  $(z, y) \in (\mathbb{R}^N)^2$ , one can obtain:

$$F(z) \leq F(y) + \langle \nabla F(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \quad (A.3)$$

Letting  $z = T_\gamma(y)$ , for all  $(x, y) \in (\mathbb{R}^N)^2$  we have:

$$\begin{aligned} F(T_\gamma(y)) - F(x) &\leq F(y) - F(x) + \langle \nabla F(y), T_\gamma(y) - y \rangle + \frac{L}{2} \|T_\gamma(y) - y\|^2 \\ &\leq \langle \nabla F(y), T_\gamma(y) - x \rangle + \frac{L}{2} \|T_\gamma(y) - y\|^2 \\ &= \frac{1}{\gamma} \langle y - T_\gamma(y), T_\gamma(y) - x \rangle + \frac{L}{2} \|T_\gamma(y) - y\|^2 \\ &= \frac{1}{2\gamma} (\|y - x\|^2 - \|T_\gamma(y) - x\|^2) + \left( \frac{L}{2} - \frac{1}{2\gamma} \right) \|T_\gamma(y) - y\|^2 \end{aligned} \quad (A.4)$$

where in the second inequality we used the fact that  $F$  is a convex function, in the first equality the definition of the operator  $T_\gamma$  ( $\gamma \nabla F(y) = y - T_\gamma(y)$ ) and in the second equality Pythagoras identity:

$$\langle y - T_\gamma(y), T_\gamma(y) - x \rangle = \frac{1}{2} (\|y - x\|^2 - \|T_\gamma(y) - x\|^2 - \|T_\gamma(y) - y\|^2) \quad (A.5)$$

By multiplying relation (A.4) by  $2\gamma$  we obtain (A.1).

The proof of the second point is similar to the first one. In particular as before, for all  $y \in \mathbb{R}^N$  and  $x^* \in X^*$  we have:

$$F(T_\gamma(y)) - F(x^*) \leq F(y) - F(x^*) + \langle \nabla F(y), T_\gamma(y) - y \rangle + \frac{L}{2} \|T_\gamma(y) - y\|^2 \quad (\text{A.6})$$

By using hypothesis  $H(\beta)$  we obtain:

$$F(T_\gamma(y)) - F(x^*) \leq \frac{1}{\beta} \langle \nabla F(y), y - x^* \rangle + \langle \nabla F(y), T_\gamma(y) - y \rangle + \frac{L}{2} \|T_\gamma(y) - y\|^2 \quad (\text{A.7})$$

By using that  $\gamma \nabla F(y) = y - T_\gamma(y)$  and Pythagoras identity, we have:

$$\begin{aligned} F(T_\gamma(y)) - F(x^*) &\leq \frac{1}{\beta\gamma} \langle y - T_\gamma(y), T_\gamma(y) - x^* \rangle + \left(\frac{L}{2} - \frac{1}{\gamma}\right) \|T_\gamma(y) - y\|^2 \\ &= \frac{1}{2\beta\gamma} (\|y - x^*\|^2 - \|T_\gamma(y) - x^*\|^2) + \left(\frac{L}{2} - \frac{1}{\gamma} + \frac{1}{2\beta\gamma}\right) \|T_\gamma(y) - y\|^2 \end{aligned} \quad (\text{A.8})$$

By multiplying the last inequality by  $2\gamma$ , we conclude the second point (A.2) of Lemma A.1.  $\square$

*Remark 4.* By choosing  $\gamma \leq \frac{1}{L}$  in Lemma A.1, it is direct that from relations (A.1) and (A.2) we obtain (respectively):

$$2\gamma(F(T_\gamma(y)) - F(x)) \leq \|y - x\|^2 - \|T_\gamma(y) - x\|^2, \forall (x, y) \in (\mathbb{R}^N)^2 \quad (\text{A.9})$$

$$2\gamma(F(T_\gamma(y)) - F(x^*)) \leq \frac{1}{\beta} \left( \|y - x^*\|^2 - \|T_\gamma(y) - x^*\|^2 \right), \forall y \in \mathbb{R}^N \text{ and } x^* \in X^* \quad (\text{A.10})$$

In particular, we have the following useful Lemma concerning the sequence generated by Algorithm 1.5.

**Lemma A.2.** *Let  $\gamma > 0$  and  $F$  satisfying  $H(\beta)$  with  $\beta \geq 1$  and  $x^* \in \arg \min F$ . Let also  $\{x_n\}_{n \geq 1}$  be the sequence generated by the (i-GD) algorithm. Then the energy-sequence  $U_n = F(x_n) - F(x^*) + \frac{\|x_n - x_{n-1}\|^2}{2\gamma}$  is non-increasing.*

*Proof.* It suffices to apply relation (A.1) of Lemma A.1, with  $\gamma \leq \frac{1}{L}$ ,  $y = y_n$  and  $x = x_n$ , in order to find:

$$F(x_{n+1}) - F(x_n) \leq \alpha_n^2 \|x_n - x_{n-1}\|^2 - \|x_{n+1} - x_n\|^2 \quad (\text{A.11})$$

By adding and subtracting  $F(x^*)$  in the left side (A.11) and rearranging the terms we find:

$$F(x_{n+1}) - F(x^*) + \frac{\|x_{n+1} - x_n\|^2}{2\gamma} \leq F(x_n) - F(x^*) + \frac{\|x_n - x_{n-1}\|^2}{2\gamma} - (1 - \alpha_n^2) \frac{\|x_n - x_{n-1}\|^2}{2\gamma} \quad (\text{A.12})$$

Since  $\alpha_n = \frac{n}{n+b} \leq 1$ , for all  $n \geq 1$ , from (A.12), we deduce that  $U_{n+1} \leq U_n$ , which concludes the proof.  $\square$

The next Lemma is a discretized version of Gronwall's Lemma ( see for example Theorem 4 in [24] or Lemma 1 in [43] ).

**Lemma A.3.** *Let  $C_0$  a positive real number and  $\{u_n\}_{n \in \mathbb{N}}$ ,  $\{v_n\}_{n \in \mathbb{N}}$  and  $\{a_n\}_{n \in \mathbb{N}}$  three non-negative sequences such that  $a_n \neq 0$  for all  $n \geq 1$  and:*

$$u_{n+1} \leq a_n u_n + v_n \quad (\text{A.13})$$

Then for all  $n \geq 1$  it holds:

$$u_{n+1} \leq \prod_{i=1}^n a_i \left( u_1 + \sum_{i=1}^n \frac{v_i}{\prod_{m=1}^i a_m} \right) \quad (\text{A.14})$$

**Lemma A.4.** Let  $a, c$  and  $C_0$  be some real numbers such that  $C_0 > 0$  and  $a > 0$  and  $\{u_n\}_{n \in \mathbb{N}}$  be a sequence of real numbers, and  $n_0 \in \mathbb{N}^*$  such that  $1 + \frac{c}{n} + \frac{a}{n^2} > 0$  for all  $n \geq n_0$ . Suppose also that for all  $n \geq n_0$ , it holds:

$$u_{n+1} \leq C_0 \prod_{i=n_0}^n \left(1 + \frac{c}{i} + \frac{a}{i^2}\right)$$

Then there exists a positive constant  $C$  and an integer  $n'_0 \geq n_0$ , such that for all  $n \geq n'_0$ , it holds:

$$u_{n+1} \leq Cn^c$$

*Proof.* In fact for all  $n \geq n_0$  we have:

$$\prod_{i=n_0}^n \left(1 + \frac{c}{i} + \frac{a}{i^2}\right) = e^{\left(\sum_{i=n_0}^n \log\left(1 + \frac{c}{i} + \frac{a}{i^2}\right)\right)} \quad (\text{A.15})$$

By using the basic inequality  $\frac{x}{1+x} \leq \log(1+x) \leq x$  for all  $x > -1$  and the summation-integral comparison test, we have from the one side:

$$\sum_{i=n_0}^n \log\left(1 + \frac{c}{i} + \frac{a}{i^2}\right) \leq \sum_{i=n_0}^n \left(\frac{c}{i} + \frac{a}{i^2}\right) \leq A + \sum_{i=n_0}^n \frac{c}{i} \leq A + c \log n \quad (\text{A.16})$$

where  $A > 0$  is a (renamed at each step) suitable positive constant.

From the other side:

$$\sum_{i=n_0}^n \log\left(1 + \frac{c}{i} + \frac{a}{i^2}\right) \geq \sum_{i=n_0}^n \left(\frac{\frac{c}{i} + \frac{a}{i^2}}{1 + \frac{c}{i} + \frac{a}{i^2}}\right) \geq \sum_{i=n_0}^n \left(\frac{c}{i+c}\right) \geq A' + c \log(n+c) \geq A' + c \log n \quad (\text{A.17})$$

where  $A' > 0$  is a (renamed at each step) suitable positive constant.

By (A.16) and (A.17) we infer that there exist  $n'_0 \in \mathbb{N}$  (such that  $n'_0 \geq n_0$ ) and some suitable positive constants  $C_1$  and  $C_2$  such that for all  $n \geq n_0$  it holds:

$$C_1 n^c \leq \prod_{i=n_0}^n \left(1 + \frac{c}{i} + \frac{a}{i^2}\right) \leq C_2 n^c \quad (\text{A.18})$$

From the hypothesis we have:

$$u_{n+1} \leq C_0 \prod_{i=n_0}^n \left(1 + \frac{c}{i} + \frac{a}{i^2}\right) \stackrel{(\text{A.18})}{\leq} Cn^c \quad (\text{A.19})$$

which concludes the proof of Lemma A.4 for a suitable positive constant  $C > 0$ . □

**Lemma A.5.** Let  $C > 0$  a positive real number,  $\alpha \in (0, 1)$  and  $\{u_n\}_{n \in \mathbb{N}}, \{z_n\}_{n \in \mathbb{N}}$  two non-negative sequences, such that for all  $n \in \mathbb{N}^*$  it holds

$$u_n \leq C + z_n u_n^\alpha \quad (\text{A.20})$$

Then for all  $n \in \mathbb{N}^*$  it holds:

$$u_n \leq 2 \max\{C, (2^\alpha z_n)^{\frac{1}{1-\alpha}}\} \quad (\text{A.21})$$

*Proof.* Let  $n \in \mathbb{N}^*$ . We split the proof in two cases:

- Firstly we suppose that  $u_n \geq (2z_n)^{\frac{1}{1-\alpha}}$ .

Since  $u_n \geq (2z_n)^{\frac{1}{1-\alpha}}$ , we have that  $1 - z_n u_n^{\alpha-1} \geq \frac{1}{2}$ , hence by using relation (A.20), we find:

$$\frac{1}{2} u_n \leq u_n (1 - z_n u_n^{\alpha-1}) \leq C$$

so that  $u_n \leq 2C$ .

- If  $u_n \leq (2z_n)^{\frac{1}{1-\alpha}}$  the result holds trivially.

□



## B Proofs of Lemmas in Sections 4.2 and 4.3

In this section we present the detailed proofs of Lemmas 4.1, 4.2 and 4.3 used in sections 4.2 and 4.3.

**Proof of Lemma 4.1.** For this proof we will frequently make use of the following basic identity:

$$\|u - z\|^2 - \|v - z\|^2 = \|u - v\|^2 + 2\langle u - v, v - z \rangle \quad \forall u, v, z \in \mathbb{R}^N \quad (\text{B.1})$$

Firstly, by applying (A.10) of Lemma A.1 with  $\gamma \leq \frac{1}{L}$  and  $y = y_n$ ,  $x = x^*$  we obtain:

$$2\gamma(F(x_{n+1}) - F(x^*)) \leq \frac{1}{\beta} \left( \|x_n + \alpha_n(x_n - x_{n-1}) - x^*\|^2 - \|x_{n+1} - x^*\|^2 \right) \quad (\text{B.2})$$

which by multiplying by  $\lambda\beta t_{n+1} > 0$ , developing the term  $\|x_n + \alpha_n(x_n - x_{n-1}) - x^*\|^2$ , is equivalent to:

$$2\gamma\lambda\beta t_{n+1}w_{n+1} \leq \lambda t_{n+1} (\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2) + 2\lambda n \langle x_n - x_{n-1}, x_n - x^* \rangle + \lambda t_{n+1} \alpha_n \|x_n - x_{n-1}\|^2 \quad (\text{B.3})$$

Since  $\alpha_n = \frac{n}{n+b} = \frac{n}{t_{n+1}}$ , by using the definitions of  $w_n$ ,  $\delta_n$  and  $h_n$ , we find:

$$\begin{aligned} 2\gamma\lambda\beta t_{n+1}w_{n+1} &\leq \lambda t_{n+1} (h_n - h_{n+1}) + 2\lambda n \langle x_n - x_{n-1}, x_n - x^* \rangle + \frac{\lambda n^2}{t_{n+1}} \delta_n \\ (\text{B.1}) \quad &= -\lambda t_{n+1} \delta_{n+1} + \frac{\lambda n^2}{t_{n+1}} \delta_n \\ &\quad + 2\lambda n \langle x_n - x_{n-1}, x_n - x^* \rangle - 2\lambda t_{n+1} \langle x_{n+1} - x_n, x_n - x^* \rangle \end{aligned} \quad (\text{B.4})$$

On the other hand, by applying (A.1) of Lemma A.1, with  $\gamma \leq \frac{1}{L}$  and  $y = y_n$ ,  $x = x_n$  we obtain:

$$2\gamma(F(x_{n+1}) - F(x_n)) \leq \alpha_n^2 \|x_n - x_{n-1}\|^2 - \|x_{n+1} - x_n\|^2 \quad (\text{B.5})$$

By adding and subtracting  $F(x^*)$  on the left side and multiplying by  $t_{n+1}^2$  on both sides, we find:

$$2\gamma t_{n+1}^2 (w_{n+1} - w_n) \leq n^2 \delta_n - t_{n+1}^2 \delta_{n+1} \quad (\text{B.6})$$

By adding relation (B.4) to relation (B.6), we obtain:

$$\begin{aligned} 2\gamma((t_{n+1}^2 + \lambda\beta t_{n+1})w_{n+1} - t_{n+1}^2 w_n) &\leq -(t_{n+1}^2 + \lambda t_{n+1})\delta_{n+1} + (n^2 + \frac{\lambda n^2}{t_{n+1}})\delta_n \\ &\quad + 2\lambda n \langle x_n - x_{n-1}, x_n - x^* \rangle - 2\lambda t_{n+1} \langle x_{n+1} - x_n, x_n - x^* \rangle \end{aligned} \quad (\text{B.7})$$

which -by adding and subtracting  $2\gamma(t_n^2 + \lambda\beta t_n)w_n$  on both sides- is equivalent to:

$$\begin{aligned} 2\gamma((t_{n+1}^2 + \lambda\beta t_{n+1})w_{n+1} - (t_n^2 + \lambda\beta t_n)w_n) &\leq 2\gamma k_{n+1} w_n - (t_{n+1}^2 + \lambda t_{n+1})\delta_{n+1} + (n^2 + \frac{\lambda n^2}{t_{n+1}})\delta_n \\ &\quad + 2\lambda n \langle x_n - x_{n-1}, x_n - x^* \rangle - 2\lambda t_{n+1} \langle x_{n+1} - x_n, x_n - x^* \rangle \\ &= 2\gamma k_{n+1} w_n - (t_{n+1}^2 + \lambda t_{n+1})\delta_{n+1} + (n^2 + 2\lambda n + \frac{\lambda n^2}{t_{n+1}})\delta_n \\ &\quad + 2\lambda n \langle x_n - x_{n-1}, x_{n-1} - x^* \rangle - 2\lambda t_{n+1} \langle x_{n+1} - x_n, x_n - x^* \rangle \end{aligned} \quad (\text{B.8})$$

where

$$\begin{aligned} k_{n+1} &= t_{n+1}^2 - \lambda\beta t_n - t_n^2 = (n+b)^2 - \lambda\beta(n+b-1) - (n+b-1)^2 \\ &= (2-\lambda\beta)(n+b-1) + 1 = (2-\lambda\beta)t_n + 1 \end{aligned} \quad (\text{B.9})$$

In addition, by developing the squares in the definition of  $v_{n+1}$  and  $v_n$ , we have:

$$\begin{aligned} v_{n+1} - v_n &= \|t_{n+1}(x_{n+1} - x_n) + \lambda(x_n - x^*)\|^2 - \|t_n(x_n - x_{n-1}) + \lambda(x_{n-1} - x^*)\|^2 \\ &= t_{n+1}^2 \|x_{n+1} - x_n\|^2 + 2\lambda t_{n+1} \langle x_{n+1} - x_n, x_n - x^* \rangle + \lambda^2 \|x_n - x^*\|^2 \\ &\quad - t_n^2 \|x_n - x_{n-1}\|^2 - 2\lambda t_n \langle x_n - x_{n-1}, x_{n-1} - x^* \rangle - \lambda^2 \|x_{n-1} - x^*\|^2 \\ &= t_{n+1}^2 \delta_{n+1} - t_n^2 \delta_n + \lambda^2 (h_n - h_{n-1}) + 2\lambda t_{n+1} \langle x_{n+1} - x_n, x_n - x^* \rangle \\ &\quad - 2\lambda t_n \langle x_n - x_{n-1}, x_{n-1} - x^* \rangle \end{aligned} \quad (\text{B.10})$$

By definition of  $E_n$  (4.12), inequality (B.8) and equality (B.10), we find:

$$\begin{aligned}
2\gamma(E_{n+1} - E_n) &= 2\gamma((t_{n+1}^2 + \lambda\beta t_{n+1})w_{n+1} - (t_n^2 + \lambda\beta t_n)w_n) + v_{n+1} - v_n \\
&\quad + \xi(h_n - h_{n-1}) + \lambda(t_{n+1}\delta_{n+1} - t_n\delta_n) \\
\text{(B.8), (B.10)} \quad &\leq 2\gamma k_{n+1}w_n + (\lambda^2 + \xi)(h_n - h_{n-1}) + (n^2 - t_n^2 + 2\lambda n - \lambda t_n + \frac{\lambda n^2}{t_{n+1}})\delta_n \\
&\quad + 2\lambda(n - t_n)\langle x_n - x_{n-1}, x_{n-1} - x^* \rangle
\end{aligned} \tag{B.11}$$

By using (B.1) we have  $h_n - h_{n-1} = \delta_n + \langle x_n - x_{n-1}, x_{n-1} - x^* \rangle$ , hence by (B.11), we find:

$$\begin{aligned}
2\gamma(E_{n+1} - E_n) &\leq 2\gamma k_{n+1}w_n + \left(n^2 - t_n^2 + 2\lambda n - \lambda t_n + \frac{\lambda n^2}{t_{n+1}} + \lambda^2 + \xi\right)\delta_n \\
&\quad + 2\left(\lambda^2 + \xi + \lambda n - \lambda t_n\right)\langle x_n - x_{n-1}, x_{n-1} - x^* \rangle
\end{aligned} \tag{B.12}$$

Since  $t_n = n + b - 1$ , by replacing  $n$  by  $t_n + 1 - b$  in (B.12) and performing some standard calculus, we find:

$$\begin{aligned}
2\gamma(E_{n+1} - E_n) &\leq 2\gamma k_{n+1}w_n + \left(2(\lambda + 1 - b)t_n + (\lambda + 1 - b)^2 + \lambda(1 - 2b) + \frac{\lambda b^2}{t_{n+1}} + \xi\right)\delta_n \\
&\quad + 2\left(\xi + \lambda(\lambda + 1 - b)\right)\langle x_n - x_{n-1}, x_{n-1} - x^* \rangle \\
(t_{n+1} \geq b) \quad &\leq 2\gamma k_{n+1}w_n + \left(2(\lambda + 1 - b)t_n + (\lambda + 1 - b)^2 + \lambda(1 - b) + \xi\right)\delta_n \\
&\quad + 2\left(\xi + \lambda(\lambda + 1 - b)\right)\langle x_n - x_{n-1}, x_{n-1} - x^* \rangle
\end{aligned} \tag{B.13}$$

By definition of  $E_n$  (4.12), we also have

$$2\gamma E_n = 2\gamma(t_n^2 + \lambda\beta t_n)w_n + (\lambda^2 + \xi)h_{n-1} + (t_n^2 + \lambda t_n)\delta_n + 2\lambda t_n \langle x_n - x_{n-1}, x_{n-1} - x^* \rangle \tag{B.14}$$

so that

$$\begin{aligned}
t_n\delta_n &= \frac{2\gamma}{t_n}E_n - 2\gamma(t_n + \lambda\beta)w_n - \frac{(\lambda^2 + \xi)}{t_n}h_{n-1} - 2\lambda\langle x_n - x_{n-1}, x_{n-1} - x^* \rangle \\
&\quad - \lambda\delta_n
\end{aligned} \tag{B.15}$$

By injecting the last equality into (B.13), we find:

$$\begin{aligned}
2\gamma(E_{n+1} - E_n) &\leq 2\gamma\frac{2(\lambda + 1 - b)}{t_n}E_n + 2\gamma(k_{n+1} - 2(\lambda + 1 - b)(t_n + \lambda\beta))w_n \\
&\quad + \left((\lambda + 1 - b)^2 + \lambda(1 - b) - 2\lambda(\lambda + 1 - b) + \xi\right)\delta_n \\
&\quad - \frac{2(\lambda + 1 - b)(\lambda^2 + \xi)}{t_n}h_{n-1} + 2\left(\xi - \lambda(\lambda + 1 - b)\right)\langle x_n - x_{n-1}, x_{n-1} - x^* \rangle
\end{aligned} \tag{B.16}$$

By choosing  $\xi = \lambda(\lambda + 1 - b)$ , in (B.16), we obtain:

$$\begin{aligned}
2\gamma(E_{n+1} - E_n) &\leq 2\gamma(k_{n+1} - 2(\lambda + 1 - b)t_n)w_n + 2\gamma\frac{2(\lambda + 1 - b)}{t_n}E_n \\
&\quad + \left((2\lambda + 1 - b)(1 - b)\right)\delta_n - \frac{2\lambda(\lambda + 1 - b)(2\lambda + 1 - b)}{t_n}h_{n-1} \\
\text{(B.9)} \quad &= 2\gamma\frac{2(\lambda + 1 - b)}{t_n}E_n + 2\gamma\left((2b - (\beta + 2)\lambda)t_n + 1 - 2\lambda\beta(\lambda + 1 - b)\right)w_n \\
&\quad + \left(((2\lambda + 1 - b)(1 - b))\right)\delta_n - \frac{2\lambda(\lambda + 1 - b)(2\lambda + 1 - b)}{t_n}h_{n-1} \\
&= 2\gamma\frac{c(\lambda)}{t_n}E_n + 2\gamma\left(A_1(\lambda)t_n + 1 - 2\lambda\beta(\lambda + 1 - b)\right)w_n + A_2(\lambda)\delta_n + \frac{A_3(\lambda)}{t_n}h_{n-1}
\end{aligned} \tag{B.17}$$

where:  $c(\lambda) = 2(\lambda + 1 - b)$ ,  $A_1(\lambda) = 2b - (\beta + 2)\lambda$ ,  $A_2(\lambda) = (2\lambda + 1 - b)(1 - b)$ , and  $A_3(\lambda) = -2\lambda(\lambda + 1 - b)(2\lambda + 1 - b)$ , which concludes the proof of the Lemma 4.1.  $\square$

**Proof of Lemma 4.2.** Firstly we suppose that  $b \leq 1 + \frac{2}{\beta}$ .

Setting  $\lambda = \frac{2b}{\beta+2} > 0$ , in the inequality (4.16) of Lemma 4.1, we find:

$$2\gamma(E_{n+1} - E_n) \leq 2\gamma \frac{c}{t_n} E_n + 2\gamma A'_1 w_n + A_2 \delta_n + \frac{A_3}{t_n} h_{n-1} \quad (\text{B.18})$$

where  $c = 2(\lambda + 1 - b) = 2 - \frac{2b\beta}{\beta+2}$  and

$$\begin{aligned} A'_1 &= 1 - 2\beta\lambda(\lambda + 1 - b) = 1 - \frac{4b\beta}{\beta+2} \left(1 - \frac{b\beta}{\beta+2}\right) \\ A_2 &= (2\lambda + 1 - b)(1 - b) = \frac{\beta-2}{\beta+2} b^2 - \frac{2\beta}{\beta+2} b + 1 = (b-1) \left(\frac{\beta-2}{\beta+2} b - 1\right) \\ \text{and } A_3 &= -2\lambda(\lambda + 1 - b)(2\lambda + 1 - b) = -\frac{2b}{\beta+2} \left(1 - \frac{b\beta}{\beta+2}\right) \left(1 - \frac{(\beta-2)b}{\beta+2}\right) \end{aligned}$$

Here we point out that in the case where  $b \leq 1 + \frac{2}{\beta}$  the constant  $A'_1$  is non-negative while  $A_2$  may be positive or negative, and  $A_3 \leq 0$ .

Without loss of generality we can suppose that the constant  $A_2$  is positive. Denoting by  $A = \max\{A'_1, A_2\} \geq 0$ , from (B.18), we obtain:

$$2\gamma(E_{n+1} - E_n) \leq 2\gamma \frac{c}{t_n} E_n + 2\gamma A w_n + A \delta_n + \frac{A_3}{t_n} h_{n-1} \quad (\text{B.19})$$

In this point, firstly we express the term  $\delta_n$  with the aid of  $E_n$  and  $w_n$  and then we regroup the different terms.

By relation (4.15), for  $\xi = \lambda(\lambda + 1 - b)$  we find:

$$2\gamma E_n \geq 2\gamma t_n^2 w_n + \frac{t_n^2}{2} \delta_n - \lambda(b-1)h_{n-1} \quad (\text{B.20})$$

Hence we have that:

$$\delta_n \leq 4\gamma \frac{E_n}{t_n^2} - 4\gamma w_n + \frac{2\lambda(b-1)}{t_n^2} h_{n-1} \quad (\text{B.21})$$

By injecting inequality (B.21) into (B.19), for all  $n \geq 1$  we find:

$$\begin{aligned} 2\gamma(E_{n+1} - E_n) &\leq 2\gamma(A - 2A)w_n + 2\gamma \frac{c}{t_n} E_n + 2\gamma \frac{2A}{t_n^2} E_n \\ &\quad + \left(\frac{4b(b-1)A}{(\beta+2)t_n} + A_3\right) \frac{h_{n-1}}{t_n} \\ &\leq 2\gamma \frac{c}{t_n} E_n + 2\gamma \frac{2A}{t_n^2} E_n + \left(\frac{4b(b-1)A}{(\beta+2)t_n} + A_3\right) \frac{h_{n-1}}{t_n} \end{aligned} \quad (\text{B.22})$$

In this point we consider the two cases depending on the value of the parameter  $b$ .

Firstly we suppose that  $b < 1 + \frac{2}{\beta}$ . In this case  $A_3 < 0$ , therefore, for  $n \in \mathbb{N}$  large enough we have that :

$$\frac{2b(b-1)A}{(\beta+2)t_n} + A_3 \leq 0 \quad (\text{B.23})$$

Hence by (B.22) we obtain:

$$2\gamma(E_{n+1} - E_n) \leq 2\gamma \left(\frac{c}{t_n} + \frac{a}{t_n^2}\right) E_n$$

which concludes the proof of the first case ( $b < 1 + \frac{2}{\beta}$ ) of Lemma 4.2 with  $a = 2A$  and  $c = 2 - \frac{2b\beta}{\beta+2}$ .

For the second case we suppose that  $b \geq 1 + \frac{2}{\beta}$  and  $F$  satisfies  $\mathcal{L}(p)$  with  $p = 2$ .

Remark that in this case ( $b \geq 1 + \frac{2}{\beta}$ ), by letting  $\lambda = \frac{2b}{\beta+2}$ , the constant  $A_3$  may be non-negative. In fact, if  $\beta > 2$  and  $1 + \frac{2}{\beta} \leq b \leq \frac{\beta+2}{\beta-2}$ , then  $A_3 \geq 0$ . Here without loss of generality we suppose that  $A_3 \geq 0$  (the case  $A_3 \leq 0$  can be treated exactly in the same way as before in the case  $b \leq 1 + \frac{2}{\beta}$ ).

In particular, by using the inequality  $\|u - v\|^2 \leq 2\|u - z\|^2 + 2\|v - z\|^2$ , for  $u = x_{n-1}$ ,  $v = x^*$  and  $z = x_n$ , in (B.19) we find:

$$\begin{aligned} 2\gamma(E_{n+1} - E_n) &\leq 2\gamma Aw_n + 2\gamma \frac{c}{t_n} E_n + \left( A + \frac{2A_3}{t_n} \right) \|x_n - x_{n-1}\|^2 + \frac{2A_3}{t_n} \|x_n - x^*\|^2 \\ &\leq 2\gamma Aw_n + 2\gamma \frac{c}{t_n} E_n + 2A\delta_n + \frac{2A_3}{t_n} \|x_n - x^*\|^2 \end{aligned} \quad (\text{B.24})$$

By using again the inequality  $\|u - v\|^2 \leq 2\|u - z\|^2 + 2\|v - z\|^2$ , with  $u = x_{n-1}$ ,  $v = x^*$  and  $z = x_n$ , in (4.15) we find:

$$\begin{aligned} 2\gamma E_n &\geq 2\gamma(t_n^2 + \lambda\beta t_n)w_n + t_n^2 \left( \frac{1}{2} + \frac{\lambda}{t_n} - \frac{2\lambda(b-1)}{t_n^2} \right) \delta_n - 2\lambda(b-1) \|x_n - x^*\|^2 \\ &\geq 2\gamma t_n^2 w_n + \frac{t_n^2}{2} \delta_n - 2\lambda(b-1) \|x_n - x^*\|^2 \end{aligned} \quad (\text{B.25})$$

Hence for  $n \in \mathbb{N}$  large enough we have

$$\delta_n \leq 4\gamma \frac{E_n}{t_n^2} - 4\gamma w_n + \frac{4\lambda(b-1)}{t_n^2} \|x_n - x^*\|^2 \quad (\text{B.26})$$

By injecting the last inequality (B.26) into (B.24) we find:

$$\begin{aligned} 2\gamma(E_{n+1} - E_n) &\leq 2\gamma(A - 4A)w_n + 2\gamma \frac{c}{t_n} E_n + 2\gamma \frac{4A}{t_n^2} E_n \\ &\quad + 2 \left( \frac{8b(b-1)A}{(\beta+2)t_n} + A_3 \right) \frac{\|x_n - x^*\|^2}{t_n} \end{aligned} \quad (\text{B.27})$$

By using Hypothesis  $\mathcal{L}(p)$  with  $p = 2$  and the uniqueness of the minimizer in inequality (B.27), we find:

$$2\gamma(E_{n+1} - E_n) \leq \left( 2K_2^{-1} \left( \frac{8b(b-1)A}{(\beta+2)t_n^2} + \frac{A_3}{t_n} \right) - 6\gamma A \right) w_n + 2\gamma \frac{c}{t_n} E_n + 2\gamma \frac{2(\beta+2)A}{t_n^2} E_n \quad (\text{B.28})$$

Therefore, for  $n \in \mathbb{N}$  large enough we have:

$$2K_2^{-1} \left( \frac{8b(b-1)A}{(\beta+2)t_n^2} + \frac{A_3}{t_n} \right) - 6\gamma A \leq 0$$

which permits to conclude the proof of Lemma 4.2 with  $a = 4A$  and  $c = 2 - \frac{2b\beta}{\beta+2}$ .  $\square$

**Proof of the point 2 in Lemma 4.2.** From Lemma 4.2, without loss of generality we can suppose that for a suitable  $n_0 \in \mathbb{N}$ , for all  $n \geq n_0$ , we have:

$$E_{n+1} - E_n \leq \frac{a}{(n+b-1)^2} E_n + \frac{c}{(n+b-1)} E_n \quad (\text{B.29})$$

with  $a = 4A$  and  $c = 2 - \frac{2b\beta}{\beta+2}$ . Equivalently:

$$E_{n+1} \leq \left( 1 + \frac{c}{t_n} + \frac{a}{t_n^2} \right) E_n \quad (\text{B.30})$$

Hence by a recurrence argument, for all  $n \geq n_0$  we find:

$$E_n \leq E_{n_0} \prod_{i=n_0}^{n-1} \left( 1 + \frac{c}{t_i} + \frac{a}{t_i^2} \right) \quad (\text{B.31})$$

Thus, by applying Lemma A.4, we can conclude that there exists some  $n_0 \in \mathbb{N}$  and a positive constant  $C > 0$ , such that for all  $n \geq n_0$  we have:  $E_n \leq Cn^c$ , as expected.  $\square$

**Proof of Lemma 4.3.** By letting  $\lambda = \frac{2}{\beta-2}$  in (4.16) of Lemma 4.1, we find:

$$2\gamma(E_{n+1} - E_n) \leq 2\gamma\left(2\left(\frac{\beta}{\beta-2} - b\right)\right)\frac{E_n}{t_n} + 2\gamma\left(B_1 t_n + 1 - \frac{4\beta}{\beta-2}\left(\frac{\beta}{\beta-2} - b\right)\right)w_n + B_2\|x_n - x_{n-1}\|^2 + \frac{B_3}{t_n}\|x_{n-1} - x^*\|^2 \quad (\text{B.32})$$

where:

$$\begin{aligned} B_1 &= (2b - (\beta + 2)\lambda) = 2\left(b - \frac{\beta + 2}{\beta - 2}\right) \\ B_2 &= (2\lambda + 1 - b)(1 - b) = (b - 1)\left(b - \frac{\beta + 2}{\beta - 2}\right) \\ \text{and } B_3 &= -2\lambda(\lambda + 1 - b)(2\lambda + 1 - b) = -\frac{4}{\beta - 2}\left(b - \frac{\beta + 2}{\beta - 2}\right)\left(b - \frac{\beta}{\beta - 2}\right) \end{aligned}$$

By definition of  $E_n$  (4.12) we have:

$$2\gamma E_n = 2\gamma(t_n^2 + \lambda\beta t_n)w_n + v_n + \lambda t_n \delta_n + \lambda(\lambda + 1 - b)h_{n-1} \quad (\text{B.33})$$

where  $\lambda = \frac{2}{\beta-2}$ .

Hence by definition of  $B_1$  and  $B_3$  and relation (B.33), we find:

$$2\gamma B_1 t_n w_n + \frac{B_3}{t_n} h_{n-1} = 2\gamma \frac{B_1 E_n}{t_n} - 2\gamma \lambda \beta B_1 w_n - \lambda B_1 \delta_n - B_1 \frac{v_n}{t_n} \quad (\text{B.34})$$

By injecting the last inequality (B.34) into (B.32) and omitting the non-positive term  $-B_1 \frac{v_n}{t_n}$ , we find:

$$2\gamma(E_{n+1} - E_n) \leq -2\gamma \frac{d}{t_n} E_n + 2\gamma B'_1 w_n + B'_2 \delta_n \quad (\text{B.35})$$

where

$$\begin{aligned} d &= \frac{4}{\beta - 2} \quad , \quad B'_1 = 1 - \frac{4\beta}{\beta - 2}\left(\frac{\beta}{\beta - 2} - b\right) - \lambda\beta B_1 = \left(\frac{\beta + 2}{\beta - 2}\right)^2 \quad \text{and} \\ B'_2 &= B_2 - \lambda B_1 = \left(b - \frac{\beta + 2}{\beta - 2}\right)^2 \end{aligned}$$

By choosing  $B' = \max\{B'_1, B'_2\}$ , from (B.35) we infer that:

$$2\gamma(E_{n+1} - E_n) \leq -2\gamma \frac{d}{t_n} E_n + 2\gamma B' w_n + B' \delta_n \quad (\text{B.36})$$

By relation (4.15) (recall that  $\lambda = \frac{2}{\beta-2}$  and  $\xi = \lambda(\lambda + 1 - b)$ ), for  $n \in \mathbb{N}$  large enough, we have that:

$$\delta_n \leq 4\gamma \frac{E_n}{t_n^2} - 4\gamma w_n + \frac{2\lambda(b-1)}{t_n^2} h_{n-1} \quad (\text{B.37})$$

Hence by injecting (B.37) into (B.36) we obtain:

$$\begin{aligned} 2\gamma(E_{n+1} - E_n) &\leq -2\gamma \frac{d}{t_n} E_n + 2\gamma \frac{2B'}{t_n^2} E_n - 2\gamma B' w_n + \frac{4(b-1)B' h_{n-1}}{(\beta-2)t_n^2} \\ &\leq -2\gamma \frac{d}{t_n} E_n + 2\gamma \frac{C}{t_n^2} E_n + \frac{C h_{n-1}}{t_n^2} \end{aligned} \quad (\text{B.38})$$

which concludes the proof of Lemma 4.3 with  $C = \max\{2B', \frac{4(b-1)B'}{\beta-2}\} > 0$ . □

## References

- [1] Vassilis Apidopoulos, Jean-François Aujol, and Charles Dossal. Convergence rate of inertial forward-backward algorithm beyond Nesterov's rule. *Mathematical Programming*, 2018.

- [2] Vassilis Apidopoulos, Jean-François Aujol, and Charles Dossal. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case  $b \leq 3$ . *SIAM Journal on Optimization*, 28(1):551–574, 2018.
- [3] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.
- [4] Hedy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.
- [5] Hedy Attouch and Alexandre Cabot. Convergence rates of inertial forward-backward algorithms. *SIAM Journal on Optimization*, 28(1):849–874, 2018.
- [6] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2):123–175, 2018.
- [7] Hedy Attouch, Zaki Chbani, and Hassan Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case  $\alpha \leq 3$ . *arXiv preprint arXiv:1706.05671*, 2017.
- [8] Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov’s accelerated forward-backward method is actually faster than  $1/k^2$ . *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.
- [9] Jean-François Aujol and Charles Dossal. Optimal rate of convergence of an ode associated to the fast gradient descent schemes for  $b > 0$ . *submitted to Journal of Differential Equations*, 2017.
- [10] Jean François Aujol, Charles Dossal, and Aude Rondepierre. Optimal convergence rates for Nesterov acceleration. *arXiv preprint arXiv:1805.05719*, 2018.
- [11] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
- [12] Jérôme Bolte, Aris Daniilidis, and Adrian. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223 (electronic), 2006.
- [13] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
- [14] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.
- [15] Alexandre Cabot, Hans Engler, and Sébastien Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society*, 361(11):5983–6017, 2009.
- [16] Luca Calatroni and Antonin Chambolle. Backtracking strategies for accelerated descent methods with smooth composite objectives. *arXiv preprint arXiv:1709.09004*, 2017.
- [17] Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the “fast iterative shrinkage/thresholding algorithm”. *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.
- [18] Dmitriy Drusvyatskiy and Adrien Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.
- [19] Olivier Fercoq and Zheng Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. *arXiv preprint arXiv:1609.07358*, 2016.
- [20] Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *arXiv preprint arXiv:1709.02300*, 2017.

- [21] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.
- [22] Guillaume Garrigos, Lorenzo Rosasco, and Silvia Villa. Convergence of the forward-backward algorithm: Beyond the worst case with the help of geometry. *arXiv preprint arXiv:1703.09477*, 2017.
- [23] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.
- [24] John M Holte. Discrete Gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7, 2009.
- [25] Franck Iutzeler and Jérôme Malick. On the proximal gradient algorithm with alternated inertia. *Journal of Optimization Theory and Applications*, 176(3):688–710, 2018.
- [26] Alexander Y Kruger. Error bounds and hölder metric subregularity. *Set-Valued and Variational Analysis*, 23(4):705–736, 2015.
- [27] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.
- [28] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.
- [29] Mingrui Liu and Tianbao Yang. Adaptive accelerated gradient converging method under h\{o}lderian error bound condition. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 3106–3116, 2017.
- [30] Stanisław. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.
- [31] Stanisław. Łojasiewicz. Sur la géométrie semi- et sous-analytique. *Annales de l'Institut Fourier. Université de Grenoble.*, 43(5):1575–1595, 1993.
- [32] Ramzi May. Asymptotic for a second order evolution equation with convex potential and vanishing damping term. *arXiv preprint arXiv:1509.05598*, 2015.
- [33] Benoît Merlet and Morgan Pierre. Convergence to equilibrium for the backward Euler scheme and applications. *Communications on Pure & Applied Analysis*, 9(3):685–702, 2010.
- [34] Ion Necoara, Yu Nesterov, and François Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.
- [35] Arkadi S. Nemirovskii and Yurii E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–309, 1985.
- [36] Yurii Nesterov. A method of solving a convex programming problem with convergence rate  $o(1/k^2)$ . In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.
- [37] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, 2013.
- [38] Brendan O’donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.
- [39] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.
- [40] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- [41] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

- [42] Vincent Roulet and Alexandre d’Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.
- [43] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, 2011.
- [44] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d’Aspremont. Integration methods and optimization algorithms. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1109–1118. Curran Associates, Inc., 2017.
- [45] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.