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

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

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

Vassilis Apidopoulos[1] Jean-François Aujol[1] Charles Dossal[2] Aude Rondepierre[2,3]

[1]Université de Bordeaux, IMB, UMR 5251, F-33400 Talence, France.
CNRS, IMB, UMR 5251, F-33400 Talence, France.
[2] IMT, Université de Toulouse, INSA Toulouse, France.
[3] LAAS, Université de Toulouse, CNRS, Toulouse, France.

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

December 24, 2018

Contents

Abstract

In this paper we study the convergence properties of the Nesterov inertial scheme which is a specific case of inertial Gradient Descent algorithm in the context of a smooth convex minimization problem, under some additional hypothesis on the local geometry of the minimizing 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 Nesterov scheme. Our analysis shows that there are some situations when Nesterov inertial scheme 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 $F : \mathbb{R}^N \longrightarrow \mathbb{R}$, is a convex function in $\mathscr{C}^{1,1}(\mathbb{R}^N)$ 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:

$$x_{n+1} = x_n - \gamma \nabla F(x_n) \tag{1.2}$$

Algorithm (1.2) is a descent scheme (i.e. $F(x_{n+1}) \leqslant F(x_n)$) provided that $0 < \gamma < \frac{2}{L}$,and without any further hypothesis made on the function F, it provides a sequence that convergences weakly to a minimizer, as also convergence rates for the objective function $F(x_n) - F(x^*)$ of order $o\big(n^{-1}\big)$. In addition this order has been proved to be optimal (see for example [29] and [20]).

In order to "accelerate" this convergence rate, in the seminal work of Nesterov [31], the author proposed an inertial version of Gradient descent algorithm with a suitable momentum term, i.e.

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

Algorithm (1.3) accelerates the convergence rate of the objective function $F(x_n) - F(x^*)$, in the sense that it is of order $O\big(n^{-2}\big)$ asymptotically. Several extensions have been made also in the non-differentiable setting (proximal-splitting-methods), see for example [21] 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 $(\alpha_n)_{n \geqslant 1}$: for any $0 < \gamma < \frac{1}{L}$, we set $x_0 = x_1 \in \mathbb{R}^N$ and for all $n \geqslant 1$:

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

$\alpha_n \in [0,1]$ may be a constant parameter or it may depends on the iteration number. In this paper we focus on the Nesterov scheme, 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 [36],[1], [7] and [6]). We therefore rewrite explicitly here what we call Nesterov (inertial) scheme in the rest of the paper:

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

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 $b \geqslant 3$, then the convergence rate of the objective function is of order of $O\big(n^{-2}\big)$ (see for example , [6], [36] 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 $o\big(n^{-2}\big)$ (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 $O\Big(n^{-\frac{2b}{3}}\Big)$.

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 $F(x_n) - F^* = O\Big(n^{-\frac{2b}{3}}\Big)$ 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 $\beta \geqslant 1$. In particular we prove that if F is strongly convex and ∇F is Lipschitz continuous, the decay is always better than $F(x_n) - F^* = O\Big(n^{-\frac{2b}{3}}\Big)$. We also prove that the actual decay for quadratic functions is $F(x_n) - F^* = O\big(n^{-b}\big)$. 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 in [20], the authors prove that the convergence rate of the gradient descent for quadratic functions is geometric. As a consequence, the actual decay $O\big(n^{-b}\big)$ for Nesterov scheme 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 [36] and [6] where it was shown that algorithm (1.5) is related to the following differential equation:

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

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 $\gamma \leqslant \frac{1}{L}$ 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 $O\big(t^{-2}\big)$ if $b \geqslant 3$ (see [36] and [6]), while for $b > 3$ this order is actually $o\big(t^{-2}\big)$ (see [28]) 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 $O\Big(t^{-\frac{2b}{3}}\Big)$ (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 compare them with related works. 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 $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function with $X^* = \arg\min F \neq \emptyset$.*

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

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

2. *Let $p \geqslant 1$. 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:*

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

The hypothesis $H(\beta)$ already used in [15, 36, 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 $\beta \geqslant 1$, 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 $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function with $X^* = \arg\min F \neq \emptyset$. If F satisfies $H(\beta)$ for some $\beta \geqslant 1$, then:*

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

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

In other words, the hypothesis $H(\beta)$ with $\beta \geqslant 1$, 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 $\mathcal{L}(p)$ with $p \geq 1$, 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]:

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

$$d(x, X^*) \leqslant \varepsilon \Rightarrow K d(x, X^*)^r \leqslant F(x) - F^*.$$

Historically, the growth condition $\mathcal{L}(p)$ is also called r-conditioning [20] or Hölderian error bounds [14], and is closely related to the Łojasiewicz inequality [26, 27], a key tool in the mathematical analysis of continuous and discrete dynamical systems [12, 13]:

Definition 2.2. *A differentiable function $F : \mathbb{R}^N \to \mathbb{R}$ 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:*

$$\forall x \in B(x^*, \varepsilon),\ \|\nabla F(x)\| \geqslant c|F(x) - F^*|^\theta. \tag{2.2}$$

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

In the convex setting, the growth condition $\mathcal{L}(p)$, $p \geqslant 1$, is indeed equivalent to the Łojasiewicz inequality, with exponent $\theta = 1 - \frac{1}{p} \in (0, 1]$ and $c = K_p^{\frac{1}{p}}$ [4, 20]. Typical examples of functions having the Łojasiewicz property are real-analytic functions and C^1 subanalytic functions [26], or semialgebraic functions [3]. 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:

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

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 $p \geqslant 2$ satisfy the global version of the hypothesis $\mathcal{L}(p)$ [20].

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 $p' \geqslant p$.

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 $F : x \in \mathbb{R} \longrightarrow |x|^\gamma$ with $\gamma > 1$, a straightforward computation shows that F satisfies $H(\beta)$ and $\mathcal{L}(p)$ if and only if $1 \leqslant \beta \leqslant \gamma \leqslant p$. More generally:

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

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:

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

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

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

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

Lemma 2.4. *Let $F : \mathbb{R}^N \to \mathbb{R}$ 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 $p \geqslant 2$.*
2. *If F satisfies $\mathcal{L}(2)$, then F automatically satisfies $H(\beta)$ with $\beta = 1 + \frac{K_2}{2L}$ and $K_2 \leqslant \frac{L}{2}$.*

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

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

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

□

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 of $F(x_n) - F(x^*)$ 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 [32], for gradient descent [20], or more general inertial schemes such as inertial gradient descent [5] or Heavy Ball methods [24]. 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.

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

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

$$F(x_n) - F(x^*) = O\Big(n^{-\frac{2b\beta}{\beta+2}}\Big) \tag{3.1}$$

If in addition $\beta \leqslant 2$ then:

$$\|x_n - x_{n-1}\| = O\Big(n^{-\frac{b\beta}{\beta+2}}\Big) \tag{3.2}$$

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

$$F(x_n) - F(x^*) = O\big(n^{-2}\big) \tag{3.3}$$

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

$$\|x_n - x_{n-1}\| = O\big(n^{-1}\big) \tag{3.4}$$

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

$$\sum_{n=0}^{+\infty} n\big(F(x_n) - F(x^*)\big) < +\infty \quad \text{and} \quad \sum_{n=0}^{+\infty} n\|x_n - x_{n-1}\|^2 < +\infty \tag{3.5}$$

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

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

$$F(x_n) - F(x^*) = O\Big(n^{-\frac{2b\beta}{\beta+2}}\Big) \quad \text{and} \quad \|x_n - x_{n-1}\| = O\Big(n^{-\frac{b\beta}{\beta+2}}\Big) \tag{3.6}$$

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:*

$$F(x_n) - F(x^*) = o\big(n^{-2}\big) \quad \text{and} \quad \|x_n - x_{n-1}\| = o\big(n^{-1}\big) \tag{3.7}$$

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 (3.5) and of Lemma B.2 in Appendix B (these two results are key elements of the convergence proof in [17]).

We define $U_n = F(x_n) - F(x^*) + \frac{\|x_n - x_{n-1}\|^2}{2\gamma}$. From (3.5) we deduce that

$$\sum_{n\geqslant 1} nU_n < +\infty \tag{3.8}$$

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

$$\sum_{k=[\frac{n}{2}]}^{n} kU_k < \varepsilon. \tag{3.9}$$

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

$$\tilde{U}_n \times \sum_{k=[\frac{n}{2}]}^{n} k \leqslant \sum_{k=[\frac{n}{2}]}^{n} kU_k < \varepsilon. \tag{3.10}$$

and thus $\tilde{U}_n = o\left(\frac{1}{n^2}\right)$. Moreover $(U_n)_{n\geqslant 1}$ is non-increasing (see Lemma B.2 in Appendix B) and thus $U_n = \tilde{U}_n$ which concludes the proof of (3.7).

The proof of the convergence of the sequence x_n to a minimizer x^* is also based on the estimates (3.5) and it is identical to the one made in [17] (see Theorem 3) and is omitted here.

Remark 1 (**The Least Square problem**). From Theorem 3.1, the point 2 is applicable under the hypothesis on the uniqueness of the minimizer of F, x^*. In fact, 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.

Let us consider the Least square inverse problem. Given $y \in \mathbb{R}^N$ and a positive-definite bounded linear operator (matrix) $A : \mathbb{R}^N \longrightarrow \mathbb{R}^N$, we consider the function $F : \mathbb{R}^N \longrightarrow \mathbb{R}$ 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.:

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}\|Ax - y\|^2 \tag{LS}$$

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

$$\begin{aligned} y_n &= x_n + \alpha_n(x_n - x_{n-1}) \\ x_{n+1} &= y_n - \gamma A^*\big(Ay_n - y\big) \end{aligned} \tag{3.11}$$

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

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

The second Theorem considers the case of a function F with a "flat" geometry near the minimizer of F. In our framework this is the case of a function F, verifying Hypotheses $H(\beta)$ and $\mathcal{L}(p)$ with $p \geqslant \beta > 2$.

Theorem 3.2. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function with a L-Lipschitz continuous gradient for some $L > 0$. Let $0 < \gamma \leqslant \frac{1}{L}$ 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 $p \geqslant \beta > 2$, and that F has a unique minimizer. If $b \geq \frac{\beta+2}{\beta-2}$, then the following estimate hold true asymptotically:

$$F(x_n) - F(x^*) = O\Big(n^{-\frac{2p}{p-2}}\Big) \tag{3.12}$$

3.2 Comments and Relation with prior works

State of the art for Nesterov scheme: It is known since [31] in 1983 that for $b = 3$ the Nesterov scheme ensures that $F(x_n) - F(x^*) = O\left(\frac{1}{t^2}\right)$, if F is convex. Actually, the proof of Nesterov was also available for $b \geqslant 3$ 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 [33]. In [17], Chambolle and the 3rd author of the present paper reminded that any $b \geqslant 3$ 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 $F(x_n) - F(x^*) = o(\frac{1}{n^2})$ when $b > 3$. More recently, following similar results from the continuous setting, Attouch et al. [7] and Apidopoulos et al. [1] proved that $F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2b}{3}}}\right)$ when $b \leqslant 3$. In Theorem 3.1 point 1 and point 2 (i) and (ii) we show that if F satisfies $H(\beta)$ for $\beta > 1$, the decay rate is actually better than $O\left(\frac{1}{n^{\frac{2b}{3}}}\right)$ for all $b \leqslant 3$ 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 $b = 3$.

In [36] Su et al. proved that for $b \leqslant 9$, if F is strongly convex, $F(x_n) - F(x^*) = O\left(\frac{1}{n^{\frac{2b}{3}}}\right)$, this result was extended to any $b > 0$ by Attouch et al. [8] for functions having a strong minimizer. The point 2 (iii) of Theorem 3.1 gives a more accurate bound when F satisfies condition $H(\beta)$ since the power $\frac{2b\beta}{\beta+2}$ is always higher than $\frac{2b}{3}$ when $\beta > 1$. We can notice that the strong convexity or the strong minimizer condition ensure that the local condition $\mathcal{L}(2)$ is in force. Moreover we have observed in Lemma 2.4 that if the gradient of F is L Lipschitz and F satisfies $\mathcal{L}(2)$ then F satisfies $H(\beta)$ for a $\beta > 1$, which implies that for such functions the decay is always faster than $O\left(\frac{1}{n^{\frac{2b}{3}}}\right)$.

This Theorem enlightens the role of the flatness hypothesis H to get a better decay rate. For classical gradient descent, the condition $\mathcal{L}(2)$ is the key to get the best decay rate, since the function $F(x) = |x|^p$ satisfies $\mathcal{L}(p)$ by definition and this function achieves the best decay rate for function satisfying $\mathcal{L}(p)$ [29]. If we consider inertial algorithm, the sharpness is not the only key. In [10], Aujol et al. show for the associated continuous problem, that this flatness hypothesis H may ensure a better decay rate, avoiding too large oscillations around the minimizer. They show that the actual decay rate for quadratic function is $O\left(\frac{1}{t^b}\right)$ which is better than $O\left(\frac{1}{t^{\frac{2b}{3}}}\right)$ that can be obtained using only strong convexity or $\mathcal{L}(2)$ condition. It is a general comment, also valid for the second theorem: for inertial algorithms, sharpness is not always the critical assumption; to get optimal rates for Nesterov scheme, one needs to control sharpness and flatness of the function. The last point of Theorem 3.1 illustrates this remark, showing for example that the decay rate is $O\left(\frac{1}{n^b}\right)$ for quadratic functions. Since this rate is proved to be optimal, see [10], for the continuous setting, we conjecture that this decay is optimal in the discrete

setting as well. One can notice that, in a continuous setting, under the sole assumption $\mathcal{L}(2)$, we cannot get a better rate than $\frac{2b}{3}$. To get the power b, a flatness hypothesis is needed.

Theorem 3.2 deals with the case when F satisfies $\mathcal{L}(p)$ with $p > 2$ but not necessarily $\mathcal{L}(2)$. As far as we know, there are no references with such hypotheses for the Nesterov scheme. This theorem ensures that if b is large enough, the decay of $F(x_n) - F(x^*)$ is faster than the rate $o\left(\frac{1}{n^2}\right)$ that can be obtained only with a convexity assumption.

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

$$x_{n+1} = x_n + \alpha_n(x_n - x_{n-1}) - \gamma\nabla F(x_n + \alpha_n(x_n - x_{n-1})) \tag{3.13}$$

or the Heavy Ball Algorithm

$$x_{n+1} = x_n + \alpha_n(x_n - x_{n-1}) - \gamma\nabla F(x_n) \tag{3.14}$$

The case $\alpha_n = 0$ of (3.13) corresponds to the classical gradient descent, the case $\alpha_n = \frac{n}{n+b}$ is the case studied in this article. If $b = 3$ we recover the original choice of Nesterov, but other choices of parameters have been studied, see for example [5].

Under the hypothesis $\mathcal{L}(2)$, the gradient descent ensures a geometrical decay.

This decay is also geometrical for the Heavy Ball method and the inertial gradient descent with fixed inertial parameter for strongly convex functions [32] (see also [33], [30]). More precisely Nesterov in 1983 [31] 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 α is the parameter of strong convexity of F, to optimize this geometrical decay. Estimating numerically $Q = \frac{L}{\alpha}$ can be a 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 [33], [18], [19], [25], [30] [16] and [34]). We can notice that for inertial methods the uniqueness of the minimizer is needed, which if not the case for gradient descent.

In the recent work [23] 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 $p \geqslant 1$.

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

Hence, even if Theorem 3.1 improves the bound for Nesterov scheme on $F(x_n) - F(x^*)$ that can be obtained with assumption $\mathcal{L}(2)$ by adding the condition $H(\gamma)$, 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, the first points of Theorem 3.2 do not assume the hypothesis $\mathcal{L}(2)$ or any strong convexity. Theorem 3.2 ensures that if $b > 1 + \frac{2}{\beta}$ the decay of $F(x^*) - F(x_n)$ is faster than $O(\frac{1}{n})$ which seems to be the best bound we can achieve with such hypotheses for Gradient descent or any inertial methods.

Moreover, the last point of 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 $o(\frac{1}{n^2})$ and better than what can be obtained with the sole hypothesis $\mathcal{L}(p)$ for gradient descent. Indeed under the assumption $\mathcal{L}(p)$, Garrigos et al. [20] proved that the gradient descent ensures a decay which is $O(n^{\frac{p}{p-2}})$ and this decay is optimal under this assumption [29].

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. In [36], 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

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla F(x(t)) = 0. \tag{3.15}$$

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 (3.15) with a constant friction term $\gamma(t) = \gamma$. Nevertheless, both algorithms do not share exactly the same properties and the same optimal tuning for parameters see [24]. 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 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.

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:

$$E_n = \varphi_n\big(F(x_n) - F(x^*)\big) + R_n \tag{4.1}$$

with φ_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 $O\big(\varphi_n^{-1}\big)$.

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 δ such that:

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

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 (4.38) for Theorem 3.2):

$$E_n \leqslant K n^{-\delta+2} \tag{4.3}$$

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

$$\mathcal{E}'(t) \leqslant \frac{c\mathcal{E}(t)}{t} + r(t) \tag{4.4}$$

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 $\mathcal{E}(t) \leqslant Kt^c$.

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

$$E_{n+1} - E_n \leqslant \frac{c}{n}E_n + r_n \tag{4.5}$$

with $c = -\delta + 2$ and a suitable sequence $\{r_n\}_{n\geqslant 1}$ 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 (4.29) 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:

$$E_n \leqslant Kn^c \tag{4.6}$$

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 ξ non positive. Thus, from (4.3), for n large enough, it follows that:
$$\varphi_n(F(x_n) - F(x^*)) - |\xi|\|x_n - x^*\|^2 \leqslant Kn^{-\delta+2} \tag{4.7}$$
and we conclude, using the growth condition $L(p)$ to bound $\|x^* - x_n\|^2$ and get inequalities such that
$$\varphi_n(F(x_n) - F(x^*)) + A_1(F(x_n) - F(x^*))^{\frac{2}{p}} \leqslant Kn^{-\delta+2} \tag{4.8}$$
which, by recalling that $\varphi_n = n^2$ and using an appropriate strategy when $p > 2$ (see Lemma B.5), leads to (4.2).

The value of δ 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 $\{E_n\}_{n\geqslant 1}$, in order to proceed to the analysis described before. Let us introduce some notations that will be useful in the coming analysis:

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

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

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

with

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

and:

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

Observe that the energy can also be expressed as:

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

Remark 2. By definition of E_n and using the convex inequality:

$$\|u\|^2 \leqslant 2\|u + v\|^2 + 2\|v\|^2 \quad , \forall u, v \in \mathbb{R}^N \tag{4.14}$$

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

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

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. *Assume that the condition $H(\beta)$ is in force with $\beta \geqslant 1$ and let $0 < \gamma \leqslant \frac{1}{L}$ and $\{x_n\}_{n\in\mathbb{N}}$ the sequence generated by Algorithm (1.5). Then for all $\lambda \geqslant 0$ and $\xi = \lambda(\lambda + b - 1)$ in the definition of E_n, the following recursive formula holds for all $n \geqslant 1$:*

$$2\gamma(E_{n+1}-E_n) \leqslant 2\gamma\frac{c(\lambda)}{t_n}E_n+2\gamma\Big(A_1(\lambda)t_n+1-2\lambda\beta(\lambda+1-b)\Big)w_n+A_2(\lambda)\|x_n-x_{n-1}\|^2+\frac{A_3(\lambda)}{t_n}\|x_{n-1}-x^*\|^2 \tag{4.16}$$

where:

$$c(\lambda) = 2(\lambda+1-b), \quad A_1(\lambda) = 2b-(\beta+2)\lambda, \quad A_2(\lambda) = (2\lambda+1-b)(1-b)$$
$$A_3(\lambda) = -2\lambda(\lambda+1-b)(2\lambda+1-b)$$

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 A.

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 $\{E_n\}_{n\geqslant 1}$ depending on the parameters β and b, and then a control estimate of E_n.

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

i. $b < 1 + \frac{2}{\beta}$

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

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

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

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

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

2. *The following estimate holds true asymptotically:*

$$E_n = O\Big(n^{2-\frac{2b\beta}{\beta+2}}\Big) \tag{4.18}$$

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 $n \geqslant n_0$, we have:

$$E_n \leqslant Ct_n^{2-\frac{2b\beta}{\beta+2}}. \tag{4.19}$$

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 ξ. 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:

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

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

In addition, if $\beta \leqslant 2$, we can get even more: combining the definition of the energy E_n as expressed in (4.13) and the estimate (4.19), we have: for all $n \geqslant n_0$,

$$\lambda t_n\big(h_n - h_{n-1}\big) + (\lambda^2 + \xi)h_{n-1} \leqslant 2\gamma E_n \leqslant C t_n^{2-\frac{2b\beta}{\beta+2}}. \tag{4.20}$$

Noticing that: $\lambda^2 + \xi = \lambda(1 + \frac{(2-\beta)b}{\beta+2}) \geqslant \lambda$ when $\beta \leqslant 2$ and that: $t_{n-1} = t_n - 1$, we then have:

$$\begin{aligned}\forall n \geqslant n_0,\ \lambda\big(t_n h_n - t_{n-1}h_{n-1}\big) &= \lambda t_n(h_n - h_{n-1}) + \lambda h_{n-1} \leqslant \lambda t_n\big(h_n - h_{n-1}\big) + (\lambda^2+\xi)h_{n-1}\\ &\leqslant C t_n^{2-\frac{2b\beta}{\beta+2}}.\end{aligned}$$

Summing over $n \geqslant n_0 + 1$ leads to:

$$\begin{aligned}\lambda t_n h_n &\leqslant \lambda t_{n_0}h_{n_0} + C\sum_{i=n_0+1}^{n} t_i^{2-\frac{2b\beta}{\beta+2}} \quad \overset{b<1+\frac{2}{\beta}}{\leqslant} \quad \lambda t_{n_0}h_{n_0} + C n t_n^{2-\frac{2b\beta}{\beta+2}}\\ &\leqslant K n^{3-\frac{2b\beta}{\beta+2}}\end{aligned} \tag{4.21}$$

asymptotically, for some suitable positive constant K. Finally, using the inequality (4.15), we have:

$$\frac{t_n^2}{2}\delta_n \leqslant 2\gamma E_n - 2\gamma t_n^2 w_n + (\lambda^2 - \xi)h_{n-1} \leqslant 2\gamma E_n + |\lambda^2 - \xi|h_{n-1}. \tag{4.22}$$

Injecting estimations (4.19) and (4.21) into (4.22) leads to: $\delta_n = O\Big(t_n^{-\frac{2b\beta}{\beta+2}}\Big)$ as expected.

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

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

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 $n \in \mathbb{N}$:

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

Using these two inequalities successively, we deduce:

$$\begin{aligned}2\gamma E_n &\geqslant 2\gamma\big(t_n^2 + \lambda\beta t_n\big)w_n + \left(\frac{1}{2} + \frac{\lambda}{t_n}\right)t_n^2\delta_n + (\xi - \lambda^2)\|x_{n-1} - x^*\|^2\\ &\geqslant 2\gamma\big(t_n^2 + \lambda\beta t_n\big)w_n + \left(\frac{1}{2} + \frac{\lambda}{t_n} - \frac{2|\xi - \lambda^2|}{t_n^2}\right)t_n^2\delta_n - 2|\xi - \lambda^2|\|x_n - x^*\|^2\\ &\geqslant 2\gamma\big(t_n^2 + \lambda\beta t_n\big)w_n + \frac{t_n^2}{4}\delta_n - 2|\xi - \lambda^2|\|x_n - x^*\|^2\end{aligned} \tag{4.23}$$

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:

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

Hence there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$, we have: $E_n \geqslant \frac{t_n^2}{2}w_n + \frac{t_n^2}{4}\delta_n$. 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 $b \geqslant 1 + \frac{2}{\beta}$. 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 different values of the parameters λ and ξ in the definition of the energy E_n. Let us set:

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

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

$$E_n = \big(t_n^2 + (b-1)\beta t_n\big)w_n + \frac{1}{2\gamma}\big(v_n + (b-1)t_n\delta_n\big) \geqslant t_n^2 w_n. \tag{4.24}$$

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 $b \geqslant 1 + \frac{2}{\beta}$, we then have:

$$\begin{aligned} \forall n \geqslant 1,\ E_{n+1} - E_n &\leqslant \big(\beta(1 + \frac{2}{\beta} - b)t_n + 1\big)w_n - \frac{1}{2\gamma}(b-1)^2\delta_n \\ &\leqslant w_n - \frac{1}{2\gamma}(b-1)^2\delta_n \leq w_n \end{aligned} \tag{4.25}$$

Injecting (4.24) into (4.25), we then obtain for all $n \geqslant 1$, $E_{n+1} \leqslant \big(1 + \frac{1}{t_n^2}\big)E_n$, which implies by a recurrence argument that:

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

By inferring Lemma B.4 with $c = 0$ and $a = 1$ we deduce that the sequence $(E_n)_{n\geqslant 1}$ is bounded. By (4.24), the sequence $(t_n^2 w_n)_{n\geqslant 1}$ is also bounded. Hence: $w_n = O\big(t_n^{-2}\big) = O\big(n^{-2}\big)$ asymptotically.

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

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

Observe now that, since $b \geqslant 1 + \frac{2}{\beta}$ and $\beta \leqslant 2$, we have: $t_n - b + 1 \leqslant t_n - 1 = t_{n-1}$, hence:

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

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

$$\forall n \geqslant n_0,\ t_n h_n - t_{n-1}h_{n-1} \leq C. \tag{4.26}$$

By summing (4.26) from n_0 to N, we obtain that for all $N \geqslant n_0$, $t_N h_N \leqslant t_{n_0}h_{n_0} + CN \leqslant t_{n_0}h_{n_0} + Ct_N$. The sequence $(h_n)_n$ is so bounded, which implies that $(x_n)_n$ is also bounded. Moreover using (4.15):

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

and the boundedness of h_n and E_n, we obtain the boundedness of the sequence $(t_n^2\delta_n)_n$, i.e. $\delta_n = O\big(n^{-2}\big)$ 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 (4.25), we have:

$$E_{n+1} - E_n \leqslant -(\beta\eta t_n - 1)w_n - \frac{1}{2\gamma}(b-1)^2\delta_n \leqslant -(\beta\eta t_n - 1)w_n \tag{4.27}$$

Moreover using the fact that there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$, we have: $\beta\frac{\eta}{2}t_n \leqslant \beta\eta t_n - 1$ and summing (4.27) over $n \in \{n_0, \cdots, N\}$, for all $N > n_0$, we find:

$$\beta\frac{\eta}{2}\sum_{n=n_0}^{N} t_n w_n \leqslant E_{n_0} < +\infty \tag{4.28}$$

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 B.1, with $\gamma \leqslant \frac{1}{L}$, $y = x_n + \alpha_n(x_n - x_{n-1})$ and $x = x_n$, or equivalently using the Lipschitz continuity of the gradient of F, we have:

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

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

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

Observe now that for all $n \geqslant 1$, we have: $2t_n - b + 1 = 2n + b - 1 \geqslant 2n$. 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.

□

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 compete proof of Theorem 3.2, we make use of the following Lemma.

Lemma 4.3. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function with a L-Lipschitz continuous gradient for some $L > 0$. Let $0 < \gamma \leqslant \frac{1}{L}$ 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 $p \geqslant \beta > 2$. 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 $b \geqslant \frac{\beta+2}{\beta-2}$, then there exist $n_0 \in \mathbb{N}$ and some constant $C > 0$ such that the following recursive formula holds true for all $n \geqslant n_0$:

$$2\gamma\big(E_{n+1} - E_n\big) \leqslant 2\gamma\bigg(-\frac{4}{\beta-2} + \frac{C}{t_n}\bigg)\frac{E_n}{t_n} + \frac{C}{t_n^2}h_{n-1}. \tag{4.29}$$

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 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:

$$E_n = O\big(n^{-m}\big), \qquad \text{with: } m = \min\Big(\frac{4}{\beta-2}, 1 + \frac{4}{p}\Big).$$

Let $d = \frac{4}{\beta-2}$. By Lemma 4.3, there exist $n_0 \in \mathbb{N}$ and a positive constant C such that

$$\forall n \geqslant n_0,\ E_{n+1} - E_n \leqslant -\frac{d}{t_n}E_n + \frac{C}{t_n^2}E_n + \frac{C}{2\gamma t_n^2}h_{n-1}. \tag{4.30}$$

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 (4.30) can be rewritten as:

$$E_{n+1} \leqslant z_n E_n + H_n \tag{4.31}$$

which, by applying Lemma B.3, implies:

$$\forall n \geqslant n_0,\ E_{n+1} \leqslant \prod_{i=n_0}^{n} z_i\bigg(E_{n_0} + \sum_{i=n_0}^{n}\frac{H_i}{\prod_{m=n_0}^{i} z_m}\bigg). \tag{4.32}$$

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

$$C_1 t_n^{-d} \leqslant \prod_{i=n_0}^{n} z_i = \prod_{i=n_0}^{n}\left(1 - \frac{d}{t_i} + \frac{C}{t_i^2}\right) \leqslant C_2 t_n^{-d}. \tag{4.33}$$

Hence by (4.32) and the definition of $H_i = \frac{C}{2\gamma t_i^2} h_{i-1}$, we find:

$$\begin{aligned} E_{n+1} &\leqslant C_2 t_n^{-d}\Big(E_{n_0} + C_1^{-1} \sum_{i=n_0}^{n} t_i^d H_i \Big) \\ &\leqslant C_2 t_n^{-d}\Big(E_{n_0} + C_3 \sum_{i=n_0}^{n} t_i^{d-2} h_{i-1} \Big) \end{aligned} \tag{4.34}$$

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:

$$h_{i-1} = \|x_{i-1} - x^*\|^2 \leqslant K_p^{-1}\big(F(x_{i-1}) - F(x^*)\big)^{\frac{2}{p}} = K_p^{-1} w_{i-1}^{\frac{2}{p}}. \tag{4.35}$$

By injecting the last inequality (4.35) into (4.34), for $n \geqslant n_0$ we find:

$$\begin{aligned} E_{n+1} &\leqslant C_2 t_n^{-d}\Big(E_{n_0} + C_3 K_p^{-1} \sum_{i=n_0}^{n} t_i^{d-2} w_{i-1}^{\frac{2}{p}} \Big) \\ &\leqslant C_2 t_n^{-d}\Big(E_{n_0} + C_3 K_p^{-1} \sum_{i=n_0}^{n} (t_i w_{i-1}) t_i^{d-1-\frac{4}{p}} \big(t_i^2 w_{i-1}\big)^{\frac{2}{p}-1} \Big) \end{aligned} \tag{4.36}$$

Moreover, since F satisfies $H(\beta)$ with $\beta > 2$, we can apply the results stated in Theorem 3.1. Here: $b \geqslant \frac{\beta+2}{\beta-2} > 1 + \frac{2}{\beta}$. Hence from relation (3.5) of Theorem 3.1, the sequence $(t_i^2 w_{i-1})_i$ is bounded and:

$$\sum_{i=n_0}^{+\infty} t_i w_{i-1} < +\infty. \tag{4.37}$$

By inferring relation (4.37), from the previous inequality (4.36), we find that for all $n \geqslant n_0$:

$$E_{n+1} \leqslant C_2 t_n^{-d}\Big(E_{n_0} + C_4 t_n^{\max\{d-1-\frac{4}{p}, 0\}} \Big),$$

hence:

$$E_{n+1} \leqslant C t_n^{-m} \tag{4.38}$$

for $m = \min\{d, 1 + \frac{4}{p}\}$ and some suitable positive constants C_4 and C.

Part 2: Once obtained the control (4.38) 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

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

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:

$$\begin{aligned} 2\gamma E_n &= 2\gamma(t_n + \lambda\beta t_n) w_n + v_n + \lambda t_n \delta_n - |\xi| \|x_{n-1} - x^*\|^2 \\ &\geqslant 2\gamma t_n^2 w_n + \frac{t_n^2}{2}\delta_n - \lambda(b-1)\|x_{n-1} - x^*\|^2. \end{aligned} \tag{4.39}$$

By using the inequality

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

in (4.39) we find:

$$2\gamma t_n^2 w_n + \Big(\frac{t_n^2}{2} - 2\lambda(b-1)\Big)\delta_n \leqslant 2\gamma E_n + 2\lambda(b-1) h_n. \tag{4.40}$$

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

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

using the control estimate (4.38) 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:

$$t_n^2 w_n \leqslant C t_n^{-m} + \frac{2\lambda(b-1)}{K_p} w_n^{\frac{2}{p}} \tag{4.41}$$

Deducing now the convergence rates on w_n is quite technical: multiplying (4.41) by t_n^m and setting $g_n = t_n^{m+2} w_n$, we find:

$$g_n \leqslant C + \frac{2\lambda(b-1)}{K} t_n^{\frac{mp-2(m+2)}{p}} g_n^{\frac{2}{p}} \tag{4.42}$$

By applying Lemma B.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:

$$\begin{aligned} g_n &\leqslant 2\max\left\{C, \left(2^{\frac{2}{p}} \frac{2\lambda(b-1)}{K_p} t_n^{\frac{mp-2(m+2)}{p}}\right)^{\frac{p}{p-2}}\right\} = 2\max\left\{C, C' t_n^{m-\frac{4}{p-2}}\right\} \\ &= O\big(t_n^M\big) \end{aligned} \tag{4.43}$$

where

$$M = \max\big\{0, m - \frac{4}{p-2}\big\} = \begin{cases} m - \frac{4}{p-2} & \text{if } p \geqslant \max(\beta, 4) \\ 0 & \text{otherwise.} \end{cases}$$

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

$$w_n = O\big(t_n^{M-m-2}\big)\,. \tag{4.44}$$

At this point we consider the different disjoint cases for the parameters (β, p) in order to precise the estimate (4.44). Recalling that $\beta \leqslant p$, $m = \min\{\frac{4}{\beta-2}, 1+\frac{4}{p}\}$ and $M = \max\{0, m - \frac{4}{p-2}\}$ we have the following results:

- If $m = \frac{4}{\beta-2}$ then (since $\beta \leqslant p$), we have necessarily that $M = m - \frac{4}{p-2}$. In that case by (4.44), 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 (4.44) we find: $w_n = O\left(t_n^{-\frac{2p}{p-2}}\right)$
  - If $M = 0$, from (4.44) we find: $w_n = O\left(t_n^{-(3+\frac{4}{p})}\right)$

The previous cases can be regrouped into two regimes $\mathcal{B}_1$ and $\mathcal{B}_2$, for the parameters $(\beta, p) \in \big\{(x,y) \in \mathbb{R}^2 :\ 2 < x \leqslant y\big\}$ with $\mathcal{B}_1$: $\big\{p \geqslant 4\big\}$ and $\mathcal{B}_2$: $\big\{p \leq 4\big\}$, such that:

- $\mathcal{B}_1$: If $p \geq 4$ then from (4.44) we obtain:

$$w_n = O\left(t_n^{-\frac{2p}{p-2}}\right) \tag{4.45}$$

- $\mathcal{B}_2$: If $p \leq 4$ then from (4.44) we obtain:

$$w_n = O\left(t_n^{-(3+\frac{4}{p})}\right) \tag{4.46}$$

In the case $p \geq 4$ (i.e. condition $\mathcal{B}_1$) we can conclude directly the proof of Theorem 3.2.

Let us now treat the case of $p \leq 4$ (i.e. condition $\mathcal{B}_2$). In this case we have $m = 1 + \frac{4}{p}$ and $M = 0$ and the estimate found in (4.46) is sub-optimal. This point is 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 (4.37) 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 (4.46) as an *a-priori* estimate, by re-injecting it in (4.35), as exactly did for (4.37). 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 $l \geqslant 1$ it holds:

$$\mu_l = 3 + \frac{2}{p}\mu_{l-1} \tag{4.47}$$

and for all $l \in \mathbb{N}$: $m_l = \min\left\{\frac{4}{\beta-2}, 1 + \frac{2}{p}\mu_l\right\}$ and $M_l = \max\left\{0, m_l - \frac{4}{p-2}\right\}$ (note that $m_0 = m$ and $M_0 = M$). Here we point out that by definition of $\{\mu_l\}_{l\geqslant 0}$, since $2 < p < 4$, the sequence μ_l diverges to infinity.

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

$$\mathcal{B}_2(l): \qquad m_l = 1 + \frac{2}{p}\mu_l \quad \text{and} \quad M_l = 0. \tag{4.48}$$

Since $\mathcal{B}_2(0)$ is in force, by relation (4.46) we have that:

$$w_n = O\left(t_n^{-(3+\frac{4}{p})}\right) = O\left(t_n^{-\mu_1}\right) \tag{4.49}$$

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

$$h_n = O\left(t_n^{-\frac{2}{p}\mu_1}\right) \tag{4.50}$$

By following the same procedure as before and injecting the inequality (4.50) into (4.34), we find:

$$E_{n+1} = O\left(t_n^{-m_1}\right) \tag{4.51}$$

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

$$w_n = O\left(t_n^{M_1-m_1-2}\right) \tag{4.52}$$

If we suppose that $\mathcal{B}_2(1)$ does not hold true (i.e. $m_1 = \frac{4}{p-2}$ or that $m_1 = 1 + \frac{2}{p}\mu_1$ and $M_1 = m_1 - \frac{4}{p-2}$), then the result of Theorem 3.2 follows directly from relation (4.52). If in the contrary $m_1 = 1 + \frac{2}{p}\mu_1$ and $M_1 = 0$, then from (4.52), it follows that:

$$w_n = O\left(t_n^{-(3+\frac{2}{p}\mu_0)}\right) = O\left(t_n^{-\mu_1}\right) \tag{4.53}$$

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

$$w_n = O\left(t_n^{M_l-m_l-2}\right) \tag{4.54}$$

In addition, if we suppose that the condition $\mathcal{B}_2(l)$, holds true for all $l \in \mathbb{N}$, then we find that for all $l \geqslant 0$ it holds:

$$w_n = O\left(t_n^{-(3+\frac{2}{p}\mu_{l-1})}\right) = O\left(t_n^{-\mu_l}\right) \tag{4.55}$$

Nevertheless, since the sequence μ_l diverges to infinity, we deduce the existence of an $l^* \in \mathbb{N}$, such that $\mathcal{B}_2(l^*)$ does not hold true (if not, from (4.48), for all $l \geqslant 0$, we would have $m_l = 1 + \frac{2}{p}\mu_l$ which is equivalent to $\mu_l \leqslant \frac{p(6-\beta)}{2(\beta-2)}$ that contradicts the fact that μ_l diverges to infinity).

Therefore (since the condition $\mathcal{B}_2(l^*)$ does not hold true), by (4.54) we deduce that:

$$w_n = O\left(t_n^{-\frac{2p}{p-2}}\right) \tag{4.56}$$

which concludes the proof of Theorem 3.2. □

A 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:

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

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

$$2\gamma\big(F(x_{n+1}) - F(x^*)\big) \leqslant \frac{1}{\beta}\bigg(\|x_n + \alpha_n(x_n - x_{n-1}) - x^*\|^2 - \|x_{n+1} - x^*\|^2\bigg) \tag{A.2}$$

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:

$$\begin{aligned} 2\gamma\lambda\beta t_{n+1} w_{n+1} \leqslant{}& \lambda t_{n+1}\big(\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2\big) + 2\lambda n\langle x_n - x_{n-1}, x_n - x^*\rangle \\ &+ \lambda t_{n+1}\alpha_n\|x_n - x_{n-1}\|^2 \end{aligned} \tag{A.3}$$

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

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

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

$$2\gamma\big(F(x_{n+1}) - F(x_n)\big) \leqslant \alpha_n^2\|x_n - x_{n-1}\|^2 - \|x_{n+1} - x_n\|^2 \tag{A.5}$$

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

$$2\gamma t_{n+1}^2\big(w_{n+1} - w_n\big) \leqslant n^2\delta_n - t_{n+1}^2\delta_{n+1} \tag{A.6}$$

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

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

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

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

where

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

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

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

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

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

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

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

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

$$
\begin{aligned}
2\gamma(E_{n+1}-E_n) &\leqslant 2\gamma k_{n+1}w_n+\bigg(2(\lambda+1-b)t_n+(\lambda+1-b)^2+\lambda(1-2b)+\frac{\lambda b^2}{t_{n+1}}+\xi\bigg)\delta_n \\
&\quad +2\bigg(\xi+\lambda(\lambda+1-b)\bigg)\langle x_n-x_{n-1},x_{n-1}-x^*\rangle \\
\big(t_{n+1}\geqslant b\big) \quad &\leqslant 2\gamma k_{n+1}w_n+\bigg(2(\lambda+1-b)t_n+(\lambda+1-b)^2+\lambda(1-b)+\xi\bigg)\delta_n \\
&\quad +2\bigg(\xi+\lambda(\lambda+1-b)\bigg)\langle x_n-x_{n-1},x_{n-1}-x^*\rangle
\end{aligned}
\tag{A.13}
$$

By definition of E_n (4.12), we also have

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

so that

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

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

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

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

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

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.

□

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

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

$$2\gamma(E_{n+1} - E_n) \leqslant 2\gamma\frac{c}{t_n}E_n + 2\gamma A_1' w_n + A_2\delta_n + \frac{A_3}{t_n}h_{n-1} \tag{A.18}$$

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

$$\begin{aligned} A_1' &= 1 - 2\beta\lambda(\lambda + 1 - b) = 1 - \frac{4b\beta}{\beta+2}\left(1 - \frac{b\beta}{\beta+2}\right) \\ A_2 &= (2\lambda + 1 - b)(1 - b) = \frac{\beta-2}{\beta+2}b^2 - \frac{2\beta}{\beta+2}b + 1 = (b-1)\left(\frac{\beta-2}{\beta+2}b - 1\right) \\ \text{and} \quad A_3 &= -2\lambda(\lambda + 1 - b)(2\lambda + 1 - b) = -\frac{2b}{\beta+2}\left(1 - \frac{b\beta}{\beta+2}\right)\left(1 - \frac{(\beta-2)b}{\beta+2}\right) \end{aligned}$$

Here we point out that in the case where $b \leqslant 1 + \frac{2}{\beta}$ the constants A_1 and A_2 are eventually positive or negative, while $A_3 \leqslant 0$.

Without loss of generality we can suppose that the constants A_1' and A_2 are positive. Denoting by $A = \max\{A_1', A_2\} \geqslant 0$, from (A.18), we obtain:

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

In this point, firstly we express the term δ_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:

$$2\gamma E_n \geqslant 2\gamma t_n^2 w_n + \frac{t_n^2}{2}\delta_n - \lambda(b-1)h_{n-1} \tag{A.20}$$

Hence we have that:

$$\delta_n \leqslant 4\gamma\frac{E_n}{t_n^2} - 4\gamma w_n + \frac{2\lambda(b-1)}{t_n^2}h_{n-1} \tag{A.21}$$

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

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

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 :

$$\frac{2b(b-1)A}{(\beta+2)t_n} + A_3 \leqslant 0 \tag{A.23}$$

Hence by (A.22) we obtain:

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

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 $b \geqslant 1 + \frac{2}{\beta}$ and F satisfies $\mathcal{L}(p)$ with $p = 2$.

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

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

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

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

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

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

$$\delta_n \leqslant 4\gamma\frac{E_n}{t_n^2} - 4\gamma w_n + \frac{4\lambda(b-1)}{t_n^2}\|x_n - x^*\|^2 \tag{A.26}$$

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

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

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

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

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

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

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

□

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 $n \geqslant n_0$, we have:

$$E_{n+1} - E_n \leqslant \frac{a}{(n+b-1)^2}E_n + \frac{c}{(n+b-1)}E_n \tag{A.29}$$

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

$$E_{n+1} \leqslant \left(1 + \frac{c}{t_n} + \frac{a}{t_n^2}\right)E_n \tag{A.30}$$

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

$$E_n \leqslant E_{n_0}\prod_{i=n_0}^{n-1}\left(1 + \frac{c}{t_i} + \frac{a}{t_i^2}\right) \tag{A.31}$$

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

□

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

$$\begin{aligned} 2\gamma\big(E_{n+1}-E_n\big) \leqslant 2\gamma\bigg(2\big(\frac{\beta}{\beta-2}-b\big)\bigg)\frac{E_n}{t_n} + 2\gamma\bigg(B_1 t_n + 1 - \frac{4\beta}{\beta-2}\big(\frac{\beta}{\beta-2}-b\big)\bigg)w_n \\ + B_2\|x_n - x_{n-1}\|^2 + \frac{B_3}{t_n}\|x_{n-1}-x^*\|^2 \end{aligned} \tag{A.32}$$

where:

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

By definition of E_n (4.12) we have:

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

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

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

$$2\gamma B_1 t_n w_n + \frac{B_3}{t_n}h_{n-1} = 2\gamma\frac{B_1 E_n}{t_n} - 2\gamma\lambda\beta B_1 w_n - \lambda B_1\delta_n - B_1\frac{v_n}{t_n} \tag{A.34}$$

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

$$2\gamma\big(E_{n+1}-E_n\big) \leqslant -2\gamma\frac{d}{t_n}E_n + 2\gamma B_1' w_n + B_2'\delta_n \tag{A.35}$$

where

$$\begin{aligned} d = \frac{4}{\beta-2} \quad , \quad B_1' = 1 - \frac{4\beta}{\beta-2}\big(\frac{\beta}{\beta-2}-b\big) - \lambda\beta B_1 = \bigg(\frac{\beta+2}{\beta-2}\bigg)^2 \quad \text{and} \\ B_2' = B_2 - \lambda B_1 = \bigg(b-\frac{\beta+2}{\beta-2}\bigg)^2 \end{aligned}$$

By choosing $B' = \max\big\{B_1', B_2'\big\}$, from (A.35) we infer that:

$$2\gamma\big(E_{n+1}-E_n\big) \leqslant -2\gamma\frac{d}{t_n}E_n + 2\gamma B' w_n + B'\delta_n \tag{A.36}$$

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:

$$\delta_n \leqslant 4\gamma\frac{E_n}{t_n^2} - 4\gamma w_n + \frac{2\lambda(b-1)}{t_n^2}h_{n-1} \tag{A.37}$$

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

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

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

□

B General Lemmas

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

First we give a basic

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

$$2\gamma(F(T_\gamma(y)) - F(x)) \leqslant \|y - x\|^2 - \|T_\gamma(y) - x\|^2 + \big(\gamma L - 1\big)\|T_\gamma(y) - y\|^2 \tag{B.1}$$

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

$$2\gamma\big(F(T_\gamma(y)) - F(x^*)\big) \leqslant \frac{1}{\beta}\bigg(\|y - x^*\|^2 - \|T_\gamma(y) - x^*\|^2\bigg) + \big(\gamma L + \frac{1}{\beta} - 2\big)\|T_\gamma(y) - y\|^2 \tag{B.2}$$

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

$$F(z) \leqslant F(y) + \langle \nabla F(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 \tag{B.3}$$

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

$$\begin{aligned} F(T_\gamma(y)) - F(x) &\leqslant F(y) - F(x) + \langle \nabla F(y), T_\gamma(y) - y\rangle + \frac{L}{2}\|T_\gamma(y) - y\|^2 \\ &\leqslant \langle \nabla F(y), T_\gamma(y) - x\rangle + \frac{L}{2}\|T_\gamma(y) - y\|^2 \\ &= \frac{1}{\gamma}\langle y - T_\gamma(y), T_\gamma(y) - x\rangle + \frac{L}{2}\|T_\gamma(y) - y\|^2 \\ &= \frac{1}{2\gamma}\big(\|y - x\|^2 - \|T_\gamma(y) - x\|^2\big) + \big(\frac{L}{2} - \frac{1}{2\gamma}\big)\|T_\gamma(y) - y\|^2 \end{aligned} \tag{B.4}$$

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_γ ($\gamma\nabla F(y) = y - T_\gamma(y)$) and in the second equality Pythagoras identity:

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

By multiplying relation (B.4) by 2γ we obtain (B.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:

$$F(T_\gamma(y)) - F(x^*) \leqslant F(y) - F(x^*) + \langle \nabla F(y), T_\gamma(y) - y\rangle + \frac{L}{2}\|T_\gamma(y) - y\|^2 \tag{B.6}$$

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

$$F(T_\gamma(y)) - F(x^*) \leqslant \frac{1}{\beta}\langle \nabla F(y), x^* - y\rangle + \langle \nabla F(y), T_\gamma(y) - y\rangle + \frac{L}{2}\|T_\gamma(y) - y\|^2 \tag{B.7}$$

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

$$\begin{aligned} F(T_\gamma(y)) - F(x^*) &\leqslant \frac{1}{\beta\gamma}\langle y - T_\gamma(y), x^* - y\rangle + \big(\frac{L}{2} - \frac{1}{\gamma}\big)\|T_\gamma(y) - y\|^2 \\ &= \frac{1}{2\beta\gamma}\big(\|y - x^*\|^2 - \|T_\gamma(y) - x^*\|^2\big) + \big(\frac{L}{2} - \frac{1}{\gamma} + \frac{1}{2\beta\gamma}\big)\|T_\gamma(y) - y\|^2 \end{aligned} \tag{B.8}$$

By multiplying the last inequality by 2γ, we conclude the second point (B.2) of Lemma B.1.

□

Remark 3. By choosing $\gamma \leqslant \frac{1}{L}$ in Lemma B.1, it is direct that from relations (B.1) and (B.2) we obtain (respectively):

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

$$2\gamma\big(F(T_\gamma(y)) - F(x^*)\big) \leqslant \frac{1}{\beta}\bigg(\|y - x^*\|^2 - \|T_\gamma(y) - x^*\|^2\bigg) \quad , \forall y \in \mathbb{R}^N \text{ and } x^* \in X^* \tag{B.10}$$

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

Lemma B.2. *Let $\gamma > 0$ and F satisfying $H(\beta)$ with $\beta \geqslant 1$ 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 (B.1) of Lemma B.1, with $\gamma \leqslant \frac{1}{L}$, $y = y_n$ and $x = x_n$, in order to find:

$$F(x_{n+1}) - F(x_n) \leqslant \alpha_n^2 \|x_n - x_{n-1}\|^2 - \|x_{n+1} - x_n\|^2 \tag{B.11}$$

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

$$F(x_{n+1}) - F(x^*) + \frac{\|x_{n+1} - x_n\|^2}{2\gamma} \leqslant F(x_n) - F(x^*) + \frac{\|x_n - x_{n-1}\|^2}{2\gamma} - (1 - \alpha_n^2)\frac{\|x_n - x_{n-1}\|^2}{2\gamma} \tag{B.12}$$

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

The next Lemma is a discretized version of Gronwall's Lemma (see for example Theorem 4 in [22] or Lemma 1 in [35]).

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

$$u_{n+1} \leqslant a_n u_n + v_n \tag{B.13}$$

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

$$u_{n+1} \leqslant \prod_{i=1}^{n} a_i \left(u_1 + \sum_{i=1}^{n} \frac{v_i}{\prod_{m=1}^{i} a_m} \right) \tag{B.14}$$

Lemma B.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 $n \geqslant n_0$. Suppose also that for all $n \geqslant n_0$, it holds:*

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

Then there exists a positive constant C, such that for all $n \geqslant n_0$, it holds:

$$u_{n+1} \leqslant C n^c$$

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

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

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

$$\sum_{i=n_0}^{n} \log\big(1 + \frac{c}{i} + \frac{a}{i^2}\big) \leqslant \sum_{i=n_0}^{n} \left(\frac{c}{i} + \frac{a}{i^2}\right) \leqslant A + \sum_{i=n_0}^{n} \frac{c}{i} \leqslant A + c\log n \tag{B.16}$$

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

From the other side:

$$\sum_{i=n_0}^{n} \log\big(1 + \frac{c}{i} + \frac{a}{i^2}\big) \geqslant \sum_{i=n_0}^{n} \left(\frac{\frac{c}{i} + \frac{a}{i^2}}{1 + \frac{c}{i} + \frac{a}{i^2}}\right) \geqslant \sum_{i=n_0}^{n} \left(\frac{c}{i+c}\right) \geqslant A' + c\log(n+c) \geqslant A' + c\log n \tag{B.17}$$

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

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

$$C_1 n^c \leqslant \prod_{i=n_0}^{n} \left(1 + \frac{c}{i} + \frac{a}{i^2}\right) \leqslant C_2 n^c \tag{B.18}$$

From the hypothesis we have:

$$u_{n+1} \leqslant C_0 \prod_{i=n_0}^{n} \left(1 + \frac{c}{i} + \frac{a}{i^2}\right) \overset{\text{(B.18)}}{\leqslant} C n^c \tag{B.19}$$

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

□

Lemma B.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*

$$u_n \leqslant C + z_n u_n^{\alpha} \tag{B.20}$$

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

$$u_n \leqslant 2\max\{C, (2^{\alpha} z_n)^{\frac{1}{1-\alpha}}\} \tag{B.21}$$

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

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

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

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

  so that $u_n \leqslant 2C$.

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

□

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] Hédy 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 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] Olivier Fercoq and Zheng Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. *arXiv preprint arXiv:1609.07358*, 2016.

[19] Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *arXiv preprint arXiv:1709.02300*, 2017.

[20] 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.

[21] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[22] John M Holte. Discrete Gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7, 2009.

[23] Franck Iutzeler and Jérôme Malick. On the proximal gradient algorithm with alternated inertia. *Journal of Optimization Theory and Applications*, 176(3):688–710, 2018.

[24] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[25] 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.

[26] 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.

[27] Stanisław. Łojasiewicz. Sur la géométrie semi- et sous-analytique. *Annales de l'Institut Fourier. Université de Grenoble.*, 43(5):1575–1595, 1993.

[28] Ramzi May. Asymptotic for a second order evolution equation with convex potential and vanishing damping term. *arXiv preprint arXiv:1509.05598*, 2015.

[29] 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.

[30] Ion Necoara, Yu Nesterov, and Fran¸ois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

[31] 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.

[32] Yurii Nesterov. Introductory lectures on convex optimization: A basic course, 2013.

[33] Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[34] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.

[35] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, 2011.

[36] 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.