Study of the behaviour of Nesterov Accelerated Gradient in a non convex setting: the strongly quasar convex case

Julien Hermant, Jean-François Aujol, Charles Dossal, Aude Rondepierre

Study of the behaviour of Nesterov Accelerated Gradient in a non convex setting: the strongly quasar convex case

J. Hermant* J.-F. Aujol* C. Dossal† A. Rondepierre†

Abstract

We study the convergence of Nesterov Accelerated Gradient (NAG) minimization algorithm applied to a class of non convex functions called strongly quasar convex functions, which can exhibit highly non convex behaviour. We show that in the case of strongly quasar convex functions, NAG can achieve an accelerated convergence speed at the cost of a lower curvature assumption. We provide a continuous analysis through high resolution ODEs, in which negative friction may appear. Finally, we investigate connections with a weaker class of non convex functions (smooth Polyak-Łojasiewicz functions) by characterizing the gap between this class and the one of smooth strongly quasar convex functions.

Keywords: Non-convex optimization, first order algorithms, strongly quasar convex, convergence rates, geometrical properties.

1 Introduction

We are interested in the following unconstrained minimization problem:

$$\min_{x\in\mathbb{R}^d} F(x) := F^* \tag{P}$$

where $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is such that $\arg\min_{x\in\mathbb{R}^d} F(x)$ is non empty. For large scale optimization, a popular class of algorithms are the *first order algorithms*, because of the relative cheapness of the iterations. These algorithms only make use of the function and its gradient, which are more computationally tractable than the Hessian that may be used by second order algorithms. We will study a specific type of first order algorithms called Nesterov Accelerated Gradient algorithms, which are variants of the gradient descent including an inertia mechanism. In the convex setting it is well known that among first order algorithms, it allows to get an accelerated rate of convergence compared to gradient descent. In the convex case, the seminal version of Nesterov achieves a rate of convergence to the minimum F^* of $\mathcal{O}\left(\frac{1}{n^2}\right)$ [31], this bound being optimal and improving over the $\mathcal{O}\left(\frac{1}{n}\right)$ rate of the gradient descent. When F is μ-strongly convex and L-smooth (i.e. C^1 with a L-Lipschitz gradient), another version of NAG [32] leads to an analogous acceleration phenomenon as we upgrade a $(1-\frac{\mu}{L})$ linear convergence rate into $(1-\sqrt{\frac{\mu}{L}})$, where $\frac{\mu}{L} \leqslant 1$ may be extremely low for high dimension functions. In recent applications, the problem of minimizing *non convex* functions has become crucial, *e.g.* in the field of machine learning. However, it is also known that the lack of regularity cancels these accelerations phenomenons. For example it has been shown that gradient descent is optimal among first order methods for the (non necessary convex) class of functions with a Lipschitz gradient [10], see also [40, 18] for similar results on other classes of functions. This means that in some cases, the benefit of Nesterov accelerated gradient in term of global convergence rate can not be proved. Convexity is however non necessary to get acceleration over gradient descent: for example [9, 22] show that a modified version of NAG accelerates over gradient descent for the class of functions with Lipschitz gradient and Hessian. In the case of convexity relaxation, the *quasar convex* (originally weak-quasi

*Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France
†IMT, Univ. Toulouse, INSA Toulouse, Toulouse, France

convex [20]) class of functions has gained a rising interest [21, 39, 14, 24, 16]. These functions are defined by the following inequality:

$$F^* \geqslant F(x) + \frac{1}{\gamma}\langle \nabla F(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2 \tag{1}$$

with x^* a minimizer, x an arbitrary point belonging to $\mathbb{R}^d$, $\gamma \in (0, 1]$, $\mu \geqslant 0$. The case $\mu > 0$ defines *strongly quasar convex* functions, which are the main focus of our work. We argue that this class is interesting to study NAG in a non convex setting. Indeed while these functions verify specific properties (only one critical point, being the global minimizer), they still may exhibit highly non convex behaviour. Moreover, it has been empirically observed that the loss function of some neural networks has a quasar-convex like structure [43].

1.1 Related work

In the case of (strongly) quasar convex functions, there exists first order algorithms achieving accelerated rates similar to those achieved in the (strongly)-convex case. In the L-smooth and $(1, \mu)-$strongly quasar convex (1) setting, [42] uses a Runge Kutta discretization procedure of the Heavy Ball ordinary differential equation to get a $1 - (\frac{\mu}{L})^\gamma$ convergence rate, $\gamma = \frac{s+3}{2(s+1)}$ where s is such that the sth derivative is Lipschitz. This means that in this case the function is needed to be high order smooth to be close to the $1 - \sqrt{\frac{\mu}{L}}$ accelerated rate. The authors of [39] apply the *continuized acceleration framework* [13] to (γ, μ)-strongly quasar convex functions, which consists in a continuous stochastic differential equation approach leading to a stochastic version of the Nesterov accelerated Gradient. This stochastic version achieves an accelerated rate $1 - \gamma\sqrt{\frac{\mu}{L}}$ in expectation, and convergence of iterations with high probability can be deduced. Finally [19, 34] shows accelerated rates in term of number of iterations, but each iteration relies on a low dimensional sub-optimization problem to solve. In [21], the cost of this sub-problem (binary line search) is explicitly computed, and the authors show that their algorithm achieves an almost optimal rate (up to a log factor) for L smooth (γ, μ)-strongly quasar convex function in term of gradient and function evaluations.
Among recent interpretations of NAG (*e.g.* [8, 23]), an important one is the ODE framework, in which these algorithms are seen as a discretization of an ordinary differential equation [35, 38]. One can show similar convergence results for the algorithms and for the solutions of these equations, mainly via Lyapunov approaches. This continuous analog gives interesting insights: it enables physical interpretation, convergence in this setting may be proved with less technical considerations than for the discrete counterpart, and importantly the strategies of proof may be adapted when we want to transpose results in the discrete setting, aiming to show algorithms convergence. This has been extensively used in a convex setting, see *e.g.* [38, 37, 5, 3, 2, 36, 27]. However to the best of our knowledge, it has not been used yet in a quasar convex setting. In this paper, we will make a step in this direction using this framework to analyze convergence in the setting of strongly quasar convex functions.

1.2 Contributions

Although in term of gradient/functions queries, accelerated convergence has already been achieved in [21], we believe that the adaptive process used lower the understanding of the algorithm behaviour in this setting, and in particular we can hardly interpret this algorithm as a discretization of a continuous damped system. Thus we address the following question.

Can Nesterov accelerated gradient algorithm achieve acceleration for the class of strongly quasar convex functions without solving an optimization sub-problem at each iteration ?

To explore this question, we provide the following contributions:

1. We show that NAG achieves an accelerated rate of convergence for the class of smooth strongly quasar convex functions when adding an assumption over lower curvature. We extend this result to composite non differentiable functions, in which case specific difficulties appear.

2. We provide a continuous analysis of the high resolution ODE associated to NAG in the strongly quasar convex setting. We will see that the gradient correction term induces weird behaviour

for non convex functions (*e.g.* potentially negative friction), although we can show accelerated convergence rates. We highlight a divergence between continuous and discrete analogies when the considered function is too non convex.

3. We prove a more general result about minimization in non convex setting: we give a geometric necessary condition for functions belonging to a subclass of strongly quasar convex functions (Polyak-Łojasiewicz functions) to be able to achieve accelerated convergence rates with first order methods. It is done by characterizing the gap between this class and the one of strongly quasar convex functions. We also give new properties of strongly quasar convex functions, as a property of on average strong convexity behaviour. We also give a construction of a pathological non convex function to prove that smooth strongly quasar convex function are not necessarily locally convex.

Organisation of paper In section 2 we define the class of functions we study in this paper, and give a brief overview of their potential non convex behaviour. In section 3 we discuss the convergence of NAG applied to strongly quasar convex functions. In section 4 we present a continuous analysis of NAG in strongly quasar convex setting. In Section 5 we present new properties of strongly quasar convex functions. In section 6 we present our numerical experiments.

2 Preliminaries

Throughout the paper, we will often consider differentiable and non necessarily convex functions $F : \mathbb{R}^d \to \mathbb{R}$. A function $F : \mathbb{R}^d \to \mathbb{R}$ is said L-smooth for some $L > 0$ if F is C^1 and has a L-Lipschitz gradient:

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d,\ \|\nabla F(x) - \nabla F(y)\| \leqslant L\|x - y\|.$$

Note that F is L-smooth if and only if F admits lower and upper quadratic bounds parameterized by L at every point. More precisely:

Property 1. *Let $F : \mathbb{R}^d \to \mathbb{R}$. F is L-smooth for some $L > 0$ if and only if it verifies for all x, y in $\mathbb{R}^d$:*

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

Proof. The fact that L-smooth implies the inequality (2) is just the well-known descent lemma, see for example [32]. The converse is also well known when dealing with convex functions. Less trivially, it still holds in the non convex case. To the best of our knowledge the equivalence is not proved in the literature, but a proof has been proposed online[1] and is adapted here to our context.
Assume first that F verify (2) with $L = 1$. Let $d \in \mathbb{R}^d$. Evaluation of (2) at 4 different pairs of points, we get:

$$F(y + d) - F(x) - \langle \nabla F(x), y + d - x\rangle \leqslant \frac{1}{2}\|y + d - x\|^2 \tag{3}$$

$$F(x - d) - F(y) - \langle \nabla F(y), x - d - y\rangle \leqslant \frac{1}{2}\|y + d - x\|^2 \tag{4}$$

$$-(F(y + d) - F(y) - \langle \nabla F(y), d\rangle) \leqslant \frac{1}{2}\|d\|^2 \tag{5}$$

$$-(F(x - d) - F(x) - \langle \nabla F(x), -d\rangle) \leqslant \frac{1}{2}\|d\|^2. \tag{6}$$

Adding all this inequalities yields:

$$\langle \nabla F(x) - \nabla F(y), x - y - 2d\rangle \leqslant \|y + d - x\|^2 + \|d\|^2 \tag{7}$$

Set $g := \nabla F(x) - \nabla F(y)$ and choose d such that $x - y - 2d = g$. Then $d = \frac{1}{2}(x - y - g)$ and $y + d - x = -\frac{1}{2}(x - y + g)$. This results in:

$$\|g\|^2 \leqslant \frac{1}{4}\|x - y + g\|^2 + \frac{1}{4}\|x - y - g\|^2 = \frac{1}{2}\|x - y\|^2 + \frac{1}{2}\|g\|^2 \tag{8}$$

[1] Characterization of Lipschitz derivative, `https://math.stackexchange.com/q/4264948` (version: 2021-10-01), Mathematics Stack Exchange

which implies F is with 1-Lipschitz gradient. It is straightforward to extend it to functions verifying (2) with arbitrary $L_0 \geqslant 0$, as it implies $\frac{F}{L_0}$ verify (2) with $L = 1$, inducing $\frac{\nabla F}{L_0}$ is 1-Lipschitz, thus inducing the result. □

This provides quadratic lower and upper bounds on the function, both parameterized by the constant L. However, we will later consider different parameterizations for these bounds. To do so we now introduce the class of (a, L)-*curvatured functions*.

Definition 1. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and (a, L) two real constants with $L > 0$ and $a \leqslant L$. The function F is said to be (a, L)-curvatured if it satisfies for all $x, y \in \mathbb{R}^d$:*

$$F(x) + \langle \nabla F(x), y - x \rangle + \frac{a}{2}\|x - y\|^2 \leqslant F(y) \leqslant F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \tag{9}$$

This is a generalisation of L-smoothness as we allow for different characterizations of lower curvature. In particular, observe that a $(-L, L)$-curvatured function is exactly a L-smooth function, a $(0, L)$-curvatured function is a L-smooth convex function, and a (μ, L)-curvatured function with $\mu > 0$ is a L-smooth and μ-strongly convex function.
In this paper we will also consider the subclass of C^2 functions having a ρ-Lipschitz Hessian (for some $\rho \geqslant 0$), *i.e.* functions that verify the following:

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \ \left|\left|\left|\nabla^2 F(x) - \nabla^2 F(y)\right|\right|\right| \leqslant \rho \|x - y\|. \tag{10}$$

where for the matrix norm $|||.|||$, we choose the norm induced by the euclidean norm on $\mathbb{R}^d$.

2.1 Strong convexity and the question of acceleration

In the paper, we study a relaxation of a well known property called strong convexity, which we recall below. Note that as we mainly studied the differentiable functions case, we state the definitions in this case. In Section 3.3 we will consider possibly non differentiable functions.

Definition 2. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. The function F is said μ-strongly convex for some $\mu > 0$ if it satisfies:*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \ F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \leqslant F(y). \tag{11}$$

This is a stronger hypothesis than convexity (which corresponds to the above inequality with $\mu = 0$). Strongly convex functions are not only lower bounded by linear approximations, but by quadratic approximations. In particular, these functions do verify the μ-quadratic growth hypothesis.

Definition 3. *Let $F : \mathbb{R}^d \to \mathbb{R}$ such that $X^* = \operatorname{argmin} F \neq \emptyset$ and $F^* = \min F$. The function F has μ-quadratic growth for some $\mu > 0$ if:*

$$\forall x \in \mathbb{R}^d, \ \frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*. \tag{12}$$

It is straightforward to see that under the strong convexity assumption, the function F has a unique minimizer. The quadratic growth property ensures that, around this minimizer, the function can not become flatter than a quadratic. This gives a control about how fast the gradient can vanish to zero when approaching the minimizer. It allows, for L-smooth and μ-strongly convex functions, the gradient descent method to generates a sequence $\{x_n\}_{n \in \mathbb{N}}$ that yields a linear convergence $F(x_n) - F^* \leqslant \mathcal{O}\left((1 - \frac{\mu}{L})^n\right)$. As mentioned in introduction, using other first order methods, this rate can be improved into a $\mathcal{O}\left((1 - \sqrt{\frac{\mu}{L}})^n\right)$ rate. Noticing that necessarily $\mu \leqslant L$, and that in high dimension, the ratio $\frac{\mu}{L}$ (which represents the inverse of the conditioning of the function to minimize) can be very small, this is actually a significant improvement.

Acceleration for relaxations of strong convexity In practice however few problems really suit the strong convexity hypothesis. Therefore, there is a need to replace it with weaker assumptions. μ-strong convexity provides a lower bound on the function, while L-smoothness provides an upper bound. Many relaxations of μ-strongly convex and L-smooth functions generalize this fact by defining another pair of assumptions, a lower one parameterized by a constant $\mu \geqslant 0$ and an upper one parameterized by a constant $L \geqslant 0$. In many cases, the property $\mu \leqslant L$ and the convergence rate associated to gradient descent in $\mathcal{O}\left((1-\frac{\mu}{L})^n\right)$ remains true. This enables to generalize, in these cases, the characterization of *acceleration* as the exchange of $\frac{\mu}{L}$ for $\sqrt{\frac{\mu}{L}}$. See [17] for an insightful discussion about lower and upper conditions.

An example of such relaxation is the class of aforementioned L-smooth functions having a μ-quadratic growth. However this is a too weak relaxation, since we lose almost all control over the function. In particular critical points can be non (global) minimum.

This is why we need to consider slightly stronger hypotheses, such as the class of μ-Polyak-Łojasiewicz functions (μ-PL), which is a non-convex relaxation of the class of μ-strongly functions:

Definition 4. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with $X^* = \arg\min\ F \neq \emptyset$ and $F^* = \min\ F$. The function F is μ-Polyak-Łojasiewicz (μ-PL) for some $\mu > 0$, if:*

$$\forall x \in \mathbb{R}^d,\ F(x) - F^* \leqslant \frac{1}{2\mu}\|\nabla F(x)\|^2. \tag{13}$$

Note that μ-PL functions have a μ-quadratic growth [15]. The Łojasiewicz property [29, 30] is a key tool in the mathematical analysis of continuous and discrete dynamical systems, initially introduced to prove the convergence of the trajectories for the gradient flow of analytic functions. The Polyak-Łojasiewicz property is nothing more than the global version of the Łojasiewicz property with an exponent $\frac{1}{2}$, and appears in important practical problems [28, 6].

It has been shown in [23] that gradient descent ensures, for L-smooth functions satisfying μ-PL property, a linear convergence: $F(x_n) - F^* \leqslant (1-\frac{\mu}{L})^n(F(x_0) - F^*)$. Importantly, in [40] is computed a lower bound of the number of gradient queries needed to achieve, with a first order method, a point $\hat{x}$ such that $F(\hat{x}) - F^* \leqslant \varepsilon(F(x_0) - F^*)$ for some $\varepsilon > 0$. They show that for every first order method, there exists a function such that this number of gradient queries is of the order $\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)$. This bound is achieved, up to a constant, by gradient descent. Strikingly, it induces that for these functions, the Nesterov accelerated gradient algorithms are prevented to achieve an accelerated linear convergence of the form $F(x_n) - F^* \leqslant K_1(1-\sqrt{\frac{\mu}{L}})^n(F(x_0) - F^*)$, where K_1 would have been a positive constant independent of μ and L. This implies that we need a more restricted class of functions to achieve this acceleration, and leads us to consider a stronger hypothesis, namely the strong quasar convexity, which is another relaxation of strong convexity.

2.2 A relaxation of strong convexity: strong quasar convexity

In this section we define the notion of strong and non-strong quasar convexity and then give a brief insight on why this class of functions can exhibit highly non convex behaviour. This last point will be more deeply discussed in section 5.

Definition 5. *Let $F : \mathbb{R}^d \to \mathbb{R}$ a differentiable function with a non-empty set of minimizers and $F^* = \min\ F$. Let x^* be a minimizer of F and $\gamma \in (0, 1]$, $\mu > 0$. The function F is said γ-quasar convex with respect to x^* if it satisfies*

$$\forall x \in \mathbb{R}^d,\ F^* \geqslant F(x) + \frac{1}{\gamma}\langle \nabla F(x), x^* - x\rangle, \tag{14}$$

and (γ, μ)-strongly quasar convex with respect to x^* if:

$$\forall x \in \mathbb{R}^d,\ F^* \geqslant F(x) + \frac{1}{\gamma}\langle \nabla F(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2. \tag{15}$$

We refer to any minimizer x^* at which (14) holds, as a quasar-convex point of F. The class of (strongly) quasar convex functions was first introduced in 2017 by [20] (with $\gamma > 0$) who refer to it as *weak quasi-convexity*, implicitly re-used by [38] and [4, 1] as a *flatness condition* (with $\gamma \geqslant 1$), and

revisited more recently in [21] that gave quasar convexity its name. A nice property of this class of functions is that any critical point of (strongly) quasar convex function F is a global minimizer of F.

Moreover, the set of minimizers X^* of a quasar convex function has some strong regularity: it is a star convex set i.e. there exists $x^* \in X^*$ such that:

$$\forall x \in X^*,\ \forall t \in [0,1],\ tx^* + (1-t)x \in X^*, \tag{16}$$

and is reduced to a single point for strongly quasar convex functions, see [21, Appendix D, Observations 3 and 4].

Lastly, observe that the Polyak-Łojasiewicz and the quadratic growth properties can be seen as relaxations of strong quasar convexity, as stated by the following result:

Proposition 1. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a (γ, μ)-strongly quasar convex function for some $(\gamma, \mu) \in (0,1] \times \mathbb{R}_+$ and x^* its minimizer. Let $F^* = \min F$. Then:*

1. *F is $\mu\gamma^2$-PL, i.e.*

$$\forall x \in \mathbb{R}^d,\ \frac{1}{2\gamma^2\mu}\|\nabla F(x)\|^2 \geqslant F(x) - F^*. \tag{17}$$

2. *[21, Corollary 1] F has a $\frac{\gamma\mu}{2-\gamma}$-quadratic growth, i.e*

$$\forall x \in \mathbb{R}^d,\ F(x) - F^* \geqslant \frac{\gamma\mu}{2(2-\gamma)}\|x - x^*\|^2. \tag{18}$$

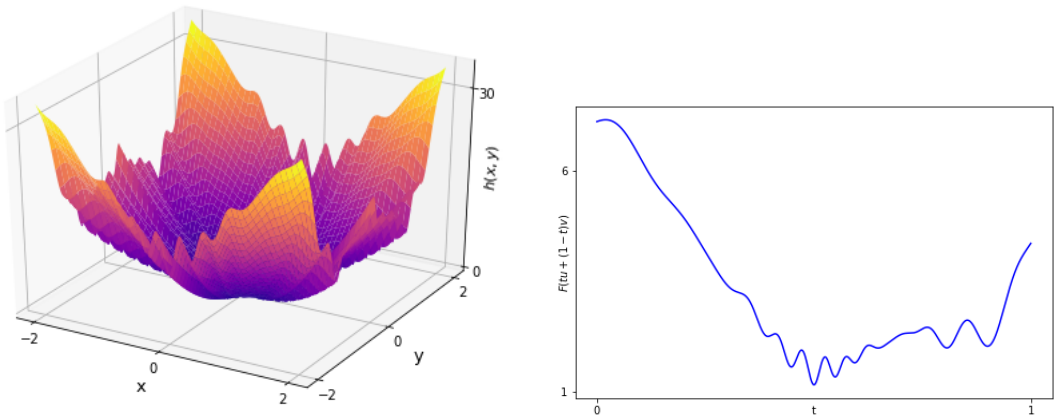

Figure 1: An example of strongly quasar convex function built as (19), whose explicit expression is given in section 6. On the left, the graph of this function. On the right, a cut of this graph along a segment $[uv]$, where $u, v \in \mathbb{R}^2$, and such that the minimizer does not belong to this segment.

Strongly quasar convex functions may exhibit highly non convex behaviour. Let us take the construction of such functions described in [25, 21] to highlight the non-convexity of strongly quasar convex functions. Let $f : \mathbb{R} \to \mathbb{R}$ such that it is (γ, μ)-strongly quasar convex, with $f^* = f(0) = 0$. Let $g : S^{d-1} \to \mathbb{R}$ be an arbitrary continuous function defined on the unit circle of $\mathbb{R}^d$ such that $g \geqslant 1$. Consider

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right),\ x \in \mathbb{R}^d. \tag{19}$$

This function is (γ, μ)-strongly quasar convex independently of the choice of g (see [21, Appendix D.3] for the non strongly quasar convex case, and see Appendix A.2 for the strongly quasar convex case). An example of such a function is displayed on the left side of Figure 1. Radially this function behaves like $cf(\|x\|)$ where c is constant. Restricted to this direction, the function is unimodal and critical points are minimizers [21, Observation 1]. However since g may be extremely non convex, we see that taking the segment between two arbitrary points x_0 and x_1, smoothness aside we will have no control

over the behaviour of the function (*e.g.* right side of Figure 1).
This lack of local regularity may not be a big deal with gradient descent, as it follows a descent direction at each iteration. However, in the case of NAG, the presence of inertia prevents from controlling the direction of the trajectory. We will see later that this lack of local regularity complicates considerably the potential acceleration of NAG.

3 Acceleration with curvature assumption for strongly quasar convex functions

3.1 The gradient descent case

As we are interested in faster algorithms than gradient descent, we first set the convergence results associated with this algorithm. We recall it is defined for some $x_0 \in \mathbb{R}^d$, $s \geqslant 0$ by the following recursive formula:

$$x_{n+1} = x_n - s\nabla F(x_n). \tag{GD}$$

Let us first recall a known result for smooth strongly convex functions. In the μ-strongly convex and L-smooth case, it is well known that gradient allows for linear decrease.

Proposition 2 ([32])**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth and μ-strongly convex function for some $0 < \mu \leqslant L$, and x^* its minimizer. Let $F^* = \min F$. Let $(x_n)_{n\in\mathbb{N}}$ be generated by (GD) with stepsize $s = \frac{1}{L}$. Then:*

$$\forall n \in \mathbb{N},\ \|x_n - x^*\|^2 \leqslant \left(1 - \frac{\mu}{L}\right)^n \|x_0 - x^*\|^2. \tag{20}$$

Proposition 2 can be extended to the class of strongly quasar convex functions. To the best of our knowledge, the following proposition is not clearly stated in literature. The case $\gamma = 1$ is proved in [17]. A result can be found in the stochastic case in [16], from which a result for the deterministic case can be deduced. Their stepsize is however lower than ours, so it yields a slower rate of convergence.

Proposition 3. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth and (γ, μ)-strongly quasar convex function for some $0 < \mu \leqslant L$, $\gamma \in (0, 1]$, and let x^* be its minimizer. Let $(x_n)_{n\in\mathbb{N}}$ be generated by (GD) with stepsize $s \leqslant \frac{1}{L}$. Then:*

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant \frac{2}{\gamma}(1 - \gamma\mu s)^n (F(x_0) - F^*). \tag{21}$$

Proof. Let x^* be the quasar convex point of F and:

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|x_n - x^*\|^2. \tag{22}$$

We compute

$$E_{n+1} - E_n = F(x_{n+1}) - F(x_n) + \frac{\mu}{2}\|x_{n+1} - x^*\|^2 - \frac{\mu}{2}\|x_n - x^*\|^2 \tag{23}$$

$$\overset{(GD)}{=} F(x_{n+1}) - F(x_n) - \mu s\langle x_n - x^*, \nabla F(x_n)\rangle + s^2\frac{\mu}{2}\|\nabla F(x_n)\|^2. \tag{24}$$

The L-smooth inequality (2) implies that: $F(x_{n+1}) - F(x_n) \leqslant -\frac{s}{2}\|\nabla F(x_n)\|^2$ provided that $s \leqslant \frac{1}{L}$. Combined with the (γ, μ)-strongly quasar convexity to control the scalar product, we get:

$$E_{n+1} - E_n \leq -\frac{s}{2}\|\nabla F(x_n)\|^2 - \gamma\mu s(F(x_n) - F^*) - \gamma s\frac{\mu^2}{2}\|x_n - x^*\|^2 + s^2\frac{\mu}{2}\|\nabla F(x_n)\|^2 \tag{25}$$

$$= \frac{s}{2}\left(\mu s - 1\right)\|\nabla F(x_n)\|^2 - \gamma\mu s\left(F(x_n) - F^* + \frac{\mu}{2}\|x_n - x^*\|^2\right) \tag{26}$$

as $s \leqslant \frac{1}{L}$ the first term is negative, inducing

$$E_{n+1} - E_n \leqslant -\gamma\mu s E_n \Rightarrow E_{n+1} \leqslant (1 - \gamma\mu s)E_n. \tag{27}$$

By induction, we then deduce:

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant (1 - \gamma\mu s)^n (F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2). \tag{28}$$

Using the $\frac{\gamma\mu}{2-\gamma}$-quadratic growth induced by the (γ, μ)-quasar strong convexity of F (see Corollary 1 [21]), we finally get the expected convergence rate. □

3.2 The Nesterov Accelerated Gradient case

Nesterov accelerated gradient scheme to optimize L-smoooth and μ-strongly convex function (NAG-SC) is often written in the following way [33, Algorithm (2.2.22)]:

$$\left\{ \begin{array}{l} y_n = x_n + \frac{1-\sqrt{\frac{\mu}{L}}}{1+\sqrt{\frac{\mu}{L}}}(x_n - x_{n-1}) \\ x_{n+1} = y_n - \frac{1}{L}\nabla F(y_n) \end{array} \right. \tag{NAG-SC 2 POINTS}$$

Introducing the auxiliary variable:

$$z_n = \left(1 + \sqrt{\frac{L}{\mu}}\right) y_n - \sqrt{\frac{L}{\mu}} x_n,$$

this algorithm can be rewritten as a 3-points scheme, see [33, Algorithm (2.2.19) with $\gamma_0 = \mu$]:

$$\left\{ \begin{array}{l} y_n = \left(\frac{1}{1+\sqrt{\frac{\mu}{L}}}\right) x_n + (1 - \frac{1}{1+\sqrt{\frac{\mu}{L}}}) z_n \\ x_{n+1} = y_n - \frac{1}{L}\nabla F(y_n) \\ z_{n+1} = (1 - \sqrt{\frac{\mu}{L}}) z_n + \sqrt{\frac{\mu}{L}}(y_n - \frac{1}{\mu}\nabla F(y_n)) \end{array} \right. \tag{NAG-SC 3 POINTS}$$

We have the following well known result.

Theorem 1 ([32])**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex function for some $0 < \mu \leqslant L$, and x^* its unique minimizer. Let $F^* = \min F$. Let $(x_n)_{n\in\mathbb{N}}$ be the sequence generated by (NAG-SC 3 POINTS) with $x_0 = z_0$. Then:*

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant \left(1 - \sqrt{\frac{\mu}{L}}\right)^n \left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right). \tag{29}$$

Observe that this is a considerable improvement over the $\mathcal{O}\left((1 - \frac{\mu}{L})^n\right)$ rate given by gradient descent (stepsize $\frac{1}{L}$), as $\frac{\mu}{L} \leqslant 1$ may be extremely low for high dimensional functions.

Let us now introduce a generalization of the classical (NAG-SC 3 POINTS) algorithm, see Algorithm 1.

Algorithm 1 Nesterov Accelerated Gradient (3 points form)

Let $z_0 = x_0$
for n = 0,. . . , **do**
  $y_n = \alpha_n x_n + (1 - \alpha_n) z_n$
  $x_{n+1} = y_n - s\nabla F(y_n)$
  $z_{n+1} = \beta_n z_n + (1 - \beta_n) y_n - \eta_n \nabla F(y_n)$
end for

For (strongly) quasar convex function minimization, this algorithm is used to get acceleration, in expectation and in probability with stochastic coefficients in [39], and deterministically in [21]. However importantly, the latter uses a binary line search at each iteration of the algorithm to compute a good α_n. The cost of this computation in term of function and gradient evaluations results in a log factor in the convergence bound. Yet we believe that this adaptive process lower the intuition we can get about the algorithm. In particular, we can hardly interpret this algorithm as a discretization of a continuous damping system. This is the gap we address in the following result, where we use the notion of *curvature function* (9):

Theorem 2. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a (γ, μ)-strongly quasar convex function for some $(\gamma, \mu) \in (0, 1] \times \mathbb{R}_+^*$ and $F^* = \min F$. Assume additionally that F is a (ρ, L)-curvatured function for some $L > 0$ and $\rho \leqslant L$. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of iterates generated by Algorithm 1 with parameters:*

$$s \leqslant \tfrac{1}{L},\ \alpha_n = \tfrac{1}{1+\sqrt{\mu s}},\ \beta_n = 1 - \gamma\sqrt{\mu s},\ \eta_n = \tfrac{\sqrt{s}}{\sqrt{\mu}}.$$

If $\rho \geqslant -\gamma\sqrt{\frac{\mu}{s}}$*, then:*

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant \frac{2}{\gamma}\left(1 - \gamma\sqrt{\mu s}\right)^n (F(x_0) - F^*). \tag{30}$$

See the proof in Appendix C.1.1, where we prove linear decrease of the following Lyapunov function

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|z_n - x^*\|^2. \tag{31}$$

More precisely, we show $E_{n+1} \leqslant (1 - \gamma\sqrt{\frac{\mu}{L}})E_n$, $\forall n \in \mathbb{N}$. The (ρ, L)-curvature assumption with $\rho < 0$ allows us to replace convexity by a weaker bound control, namely:

$$\forall n \in \mathbb{N},\ \langle \nabla F(y_n), x_n - y_n\rangle + F(y_n) - F(x_n) \leqslant -\frac{\rho}{2}\|x_n - y_n\|^2.$$

Fixing $s = \frac{1}{L}$, the curvature bound becomes $\rho \geqslant -\gamma\sqrt{\mu L} = -\gamma\sqrt{\frac{\mu}{L}}L$. Parameterizing the algorithm with an arbitrary step size $s \leqslant \frac{1}{L}$ allows us to highlight that we can trade restriction on the curvature with convergence speed. Formally:

Corollary 1. *Let* $F : \mathbb{R}^d \to \mathbb{R}$ *be a* L*-smooth and* (γ, μ)*-strongly quasar convex function for some* $0 < \mu \leqslant L$*,* $\gamma \in (0, 1]$*, and let* $F^* = \min F$*. Let* $(x_n)_{n\in\mathbb{N}}$ *generated by Algorithm 1 with parameters* $s = \gamma^2\frac{\mu}{L^2}$*,* $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$*,* $\beta_n = 1 - \gamma\sqrt{\mu s}$ *and* $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$*. Then:*

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant \frac{2}{\gamma}\left(1 - \gamma^2\frac{\mu}{L}\right)^n (F(x_0) - F^*). \tag{32}$$

Proof. Just solve $-\gamma\sqrt{\frac{\mu}{s}} = -L$, we get $s = \gamma^2\frac{\mu}{L^2}$. □

We see that we can delete the curvature assumption, but at the cost of acceleration. The negative curvature problem has already been observed under other non convex hypothesis. Authors in [9, 22] prove acceleration convergence to a critical point in a Hessian Lipschitz setting, where they use NAG when curvature between iterates is not too negative, otherwise an alternative step using Hessian Lipschitz property is performed. In our case, alternative step using Hessian Lipschitz property is hardly useful as our Lyapunov designed to obtain linear convergence is less adapted to this argument.

Comparison with the algorithm of [21, Algorithm 3] (binary search) The algorithm 3 presented in [21] is parameterized with $s = \frac{1}{L_n}$ (where the parameter L_n is computed by backtracking) and with the same sequences η_n and β_n as defined in Corollary 1, while their α_n sequence is computed using a binary line search process, and is not explicit. This results in a $\log\left(\gamma^{-1}\frac{L}{\mu}\right)$ factor added to the number of gradient and function evaluations needed, that does not appear in our convergence rate. This means that on the restricted class of functions defined in Theorem 2 with $s = \frac{1}{L}$, the bound provided in Theorem 2 is better by this log factor.

2 points scheme form of algorithm 1 Following arguments of [26], we can rewrite our algorithm in a 2 points sequence scheme, as stated with this result:

Proposition 4. *The algorithm 1 with parameters* $s \leqslant \frac{1}{L}$*,* $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$*,* $\beta_n = 1 - \gamma\sqrt{\mu s}$ *and* $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$ *can be written as the following 2 points scheme:*

$$\begin{cases} y_n = x_n + \frac{1-\gamma\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_n - x_{n-1}) + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\gamma - 1)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla F(y_n) \end{cases} \tag{NAG-SQC 2 POINTS}$$

When $\gamma = 1$, we recover the classical Nesterov scheme for minimizing strongly convex functions. Note that 2 points scheme versions are widely present in the literature. Considering the 2 points or the 3 points version gives different interpretations of the algorithm, and switching from a version to another is not obvious. The proof of Proposition 4 is in Appendix C.1.2.

3.3 Composite non differentiable case

A natural way to extend the differential case to the non differentiable one is to consider the class of composite functions $F = f + g$, where f is assumed to be at least differentiable and strongly quasar convex, and g convex, proper lower semi-continuous, and such that its proximal operator [11] can be computed:

$$\text{prox}_g(x) = \arg\min_y \left(g(y) + \frac{1}{2}\|x - y\|^2 \right). \tag{33}$$

Typical example of such functions are the composite functions with a ℓ_1 penalisation term: $F = f + \|x\|_1$. The idea is then simply to replace the gradient in Algorithm 1 by the composite gradient mapping:

$$\nabla f(y_n) \to \frac{1}{s}\left(y_n - \text{prox}_{sg}(y_n - s\nabla f(y_n))\right) \tag{34}$$

which may be seen as a generalization of the gradient (we recover it when $g \equiv 0$). It leads us to Algorithm 2.

Algorithm 2 Proximal Nesterov Accelerated Gradient (3 points form)

Soit $z_0 = x_0$
for k = 0,. . . , **do**
  $y_n = \alpha_n x_n + (1 - \alpha_n) z_n$
  $x_{n+1} = \text{prox}_{sg}(y_n - s\nabla f(y_n)) := T_s(y_n))$
  $z_{n+1} = \beta_n z_n + (1 - \beta_n) y_n - \frac{\eta_n}{s}(y_n - T_s(y_n))$
end for

Extending Theorem 2 to the non-differentiable case introduces an additional technicality since a minimizer/quasar point of f is generally not a minimizer of the composite function F. In other words, if f is strongly quasar convex with respect to its minimizer x_f^*, there is clearly no guarantee that this assumption holds at a minimizer x_F^* of F. Therefore, we propose an extension of strong quasar convexity with respect to another point than a minimizer as suggested in [21, Appendix D.2]:

Definition 6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\hat{x} \in \mathbb{R}^d$. The function f is said $(1, \mu)$-strongly quasar convex with respect to $\hat{x}$ if:*

$$\forall x \in \mathbb{R}^d,\ \forall t \in [0,1],\ f(t\hat{x} + (1-t)x) + t\,(1-t)\,\frac{\mu}{2}\|\hat{x} - x\|^2 \leqslant tf(\hat{x}) + (1-t)f(x),$$

or equivalently, when f is additionally assumed to be differentiable:

$$\forall x \in \mathbb{R}^d,\ f(\hat{x}) \geqslant f(x) + \langle \nabla f(x), \hat{x} - x\rangle + \frac{\mu}{2}\|\hat{x} - x\|^2. \tag{35}$$

The equivalence is showed in [21, Lemma 11], whose proof works for $\hat{x}$ not being a minimizer, if we consider (γ, μ) strong convexity with $\gamma = 1$.

The $\gamma < 1$ case Taking $\gamma < 1$ does not cope well with quasar convexity with respect to an arbitrary point. Assume that for some $f \in C^1$, $\mu > 0$ and $\gamma \in (0, 1]$ the following holds:

$$\forall x \in \mathbb{R}^d,\ f(\hat{x}) \geqslant f(x) + \frac{1}{\gamma}\langle \nabla f(x), \hat{x} - x\rangle + \frac{\mu}{2}\|\hat{x} - x\|^2. \tag{36}$$

with $\hat{x}$ such that $\nabla f(\hat{x}) \neq 0$. Set $x_h = \hat{x} - h\nabla f(\hat{x})$, $h > 0$. Using a order 1 Taylor-Young development we have:

$$f(\hat{x}) = f(x_h) + \langle \nabla f(x_h), \hat{x} - x_h\rangle + o(\|\hat{x} - x_h\|) = f(x_h) + h\langle \nabla f(x_h), \nabla f(\hat{x})\rangle + o(h) \tag{37}$$

Combining this with (36) evaluated in $x = x_h$, we have:

$$\langle \nabla f(x_h), \nabla f(\hat{x})\rangle \left(1 - \frac{1}{\gamma}\right) + \frac{o(h)}{h} \geqslant \frac{h\mu}{2}\|\nabla f(\hat{x})\|^2 \tag{38}$$

As h goes to 0, the right hand side goes to 0 while the left hand side goes to $\|\nabla f(\hat{x})\| \left(1 - \frac{1}{\gamma}\right)$ because f is C^1. As the latter is strictly negative for $\gamma \in (0, 1)$, it implies that in our definition γ must be equal to 1.
Observe now that Definition 6 is not an empty definition, and that such functions can easily be defined. A possible construction consists in adding some convex function to a $(1, \mu)$-strongly quasar convex function.

Lemma 1. *Let f: $\mathbb{R}^d \to \mathbb{R}$ be $(1, \mu)$-strongly quasar convex with respect to $\hat{x}$ and $g : \mathbb{R}^d \to \mathbb{R}$ be convex, both non necessarily differentiable. Then $F := f + g$ is $(1, \mu)$-strongly quasar convex with respect to $\hat{x}$.*

Proof. The proof is straightforward by summing the respective inequalities of $(1, \mu)$-strong quasar convexity of f and convexity of g written between any $x \in \mathbb{R}^d$ and $\hat{x}$. □

As a special case of Lemma 1, observe that if f is $(1, \mu)$-strongly quasar convex with respect to its minimizer x_f^*, then x_f^* is not necessarily a minimizer of F, but it is nevertheless a quasar convex point of F.

Convergence result for Algorithm 2 Using the extended definition of $(1, \mu)$-strongly quasar convexity, we are ready to state our result.

Theorem 3. *Let $F = f + g$ where $f : \mathbb{R}^d \to \mathbb{R}$ is a L-smooth function for some $L > 0$ and $g : \mathbb{R}^d \to \mathbb{R}$ is convex, proper, lower semi-continuous. Assume that F has a non empty set of minimizers and that f is $(1, \mu)$-strongly quasar convex with respect to $x_F^* \in \arg\min\ F$ for some $0 < \mu \leqslant L$ and (ρ, L)-curvatured for some $\rho \leqslant L$. Let $(x_n)_{n\in\mathbb{N}}$ be the sequence of iterates generated by Algorithm 2 with parameters:*

$$s \leqslant \tfrac{1}{L},\ \alpha_n = \tfrac{1}{1+\sqrt{\mu s}},\ \beta_n = 1 - \sqrt{\mu s},\ \eta_n = \tfrac{\sqrt{s}}{\sqrt{\mu}}.$$

If $\rho \geqslant -\sqrt{\frac{\mu}{s}}$, then:

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant 2\left(1 - \sqrt{\mu s}\right)^n \left(F(x_0) - F^*\right). \tag{39}$$

Proof is in Appendix C.1.3. It uses the same Lyapunov function as in Theorem 2 and it makes use of the Prox-Grad theorem from [7].
It might seems weird, for practical purpose, to ask strongly quasar convexity at a point which is not a minimizer of the aforementioned function. It allows to show that theoretically, convexity of f is not necessary to extend the accelerated result in the differentiable case to the composite case. Note that together with difficulties of defining (strongly) quasar convex functions with respect to an arbitrary point, arguments used in our proof do not work if $\gamma < 1$. In that case, the $\gamma < 1$ relaxation is non trivial, while in some cases it just demands computations adjusting.

Remark As a direct consequence of Lemma 1, if F is such as it is defined in Theorem 3, it is then $(1, \mu)$-strongly quasar convex with respect to its minimizer x_F^*.

4 Continuous analysis

4.1 High resolution ordinary differential equation

Considering Nesterov accelerated gradient algorithm as a discretization of an ODE is a powerful tool, helping to gain intuition, finding and generalizing new convergence results [38, 37].
In [36] are introduced high resolution ODEs to study first order optimization algorithms. The interest of these ODEs is to describe more accurately the behaviour of the corresponding algorithms than those commonly used and known as low-resolution ODEs. This is made by keeping $\mathcal{O}(\sqrt{s})$ terms during the discretization process, inducing that the ODE depends on the stepsize s. In particular, while Polyak's Heavy Ball [35] and NAG can be seen as discretizations of the same low resolution ODE:

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \nabla F(X(t)) = 0$$

these algorithms correspond to two different high resolutions ODEs (see [36] for more details):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1+\sqrt{\mu s})\nabla F(X(t)) = 0 \tag{HB-ODE}$$

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 F(X(t))\dot{X}(t) + (1+\sqrt{\mu s})\nabla F(X(t)) = 0 \tag{NAG-SC-ODE}$$

As Polyak's Heavy ball cannot achieve an accelerated rate for smooth strongly convex functions, while NAG can, understanding how the two high resolution ODEs differ is thus interesting. The only differing term is $\sqrt{s}\nabla^2 F(X(t))\dot{X}(t)$, which corresponds in the algorithm to what the authors refer to as a gradient correction term. In (NAG-SC-ODE), factorizing $\dot{X}$ terms, we see that the damping coefficient becomes $2\sqrt{\mu} + \sqrt{s}\nabla^2 F(X(t))$, which is now adaptive to the position of X. In particular if $\dot{X}$ is colinear with the eigenvector corresponding to the highest eigenvalue of $\nabla^2 F$ (possibly L), the new damping rate increases, thus reducing oscillations.
As we stated in Proposition 4, in the case of $(1,\mu)$-strongly quasar convex functions, Algorithm 1 with right parameters defines the same algorithm as the classical Nesterov Algorithm to minimize μ-strongly convex and L-smooth functions. In that case, we will get the same high resolution ODE (NAG-SC-ODE). We see then immediately that this Hessian term will induce specific behaviour in a non convex case. In particular if $\dot{X}$ is colinear to the eigenvector of $\nabla^2 F(X(t))$ associated with eigenvalue $-L$, the damping rate becomes

$$(2\sqrt{\mu} - L\sqrt{s})\dot{X}(t).$$

Take $s = \frac{1}{L}$ and it is negative (if μ is not too close to L). The case of negative damping rate is unusual. In particular, it induces the increase of the total mechanical energy $(F(X(t)) - F^* + \frac{1}{2}\|\dot{X}(t)\|^2)$.

High resolution ODE for Algorithm 1 Recall that with the choice of parameters of Theorem 2, Algorithm 1 is

$$\begin{cases} y_n = \frac{1}{1+\sqrt{\mu s}}x_n + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}z_n \\ x_{n+1} = y_n - s\nabla F(y_n) \\ z_{n+1} = 1 - \gamma\sqrt{\mu s}z_n + \gamma\sqrt{\mu s}y_n - \sqrt{\frac{s}{\mu}}\nabla F(y_n) \end{cases} \tag{NAG-SQC-DISCRETE}$$

The high resolution ODE associated with (NAG-SQC-DISCRETE) can be written in the following way:

$$\left(1 + \frac{1-\gamma}{2}\sqrt{\mu s}\right)\ddot{X}(t) + (1+\gamma)\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 F(X(t))\dot{X}(t) + (1+\gamma\sqrt{\mu s})\nabla F(X(t)) = 0 \tag{NAG-SQC-ODE}$$

The derivation of this ODE is detailed in Appendix D. Observe that if taking $\gamma = 1$, we recover (NAG-SC-ODE). Despite the eventual negative friction of (NAG-SQC-ODE), one can get results for the convergence of the solution of this equation that are similar to the one obtained in the strongly convex setting [37], up to a γ factor.

Proposition 5. *Let F be $(\gamma,\mu)-$strongly quasar convex and L-smooth for some $0 < \mu \leqslant L$, $\gamma \in (0,1]$. Assume X is solution of (NAG-SQC-ODE) with $0 < s \leqslant \frac{1}{L}$, $X(0) = X_0$ and $\dot{X}(0) = 0$. Then:*

$$F(X(t)) - F^* \leqslant K_0(\gamma,\mu,L,s)\frac{1}{\gamma}(F(X_0) - F^*)e^{-\gamma\frac{\sqrt{\mu}}{2}t} \tag{40}$$

where $K_0(\gamma,\mu,L,s)$ can be uniformly bounded by 7.

To prove Proposition 5, we aim to show the linear decrease of a Lyapunov function of the form:

$$\mathcal{E}(t) = \delta(F(X(t)) - F^*) + \frac{1}{2}\left\|\left(1 + \frac{1-\gamma}{2}\sqrt{\mu s}\right)\dot{X}(t) + \lambda(X(t) - x^*) + \sqrt{s}\nabla F(X(t))\right\|^2. \tag{41}$$

where δ, λ belong in $\mathbb{R}$ and are well chosen parameters. See Appendix D for the proof.
The $\sqrt{\mu}$ in the exponential exponent is how we characterize possibility of acceleration in the continuous case for strongly convex functions. The gradient flow (ODE version of gradient descent) achieves a

similar rate of convergence, with a μ exponent instead of $\sqrt{\mu}$.
The result of Proposition 5 is then quite surprising: it means that occurring of negative damping along the trajectory is not a problem, and that accelerated convergence occurs anyway despite this weird behaviour. Thus, while non convexity may impact negatively the convergence of Algorithm 1, it is not the case for the associated continuous system. This would indicate that non convexity impacts the discretization process. We confirm this intuition in the next section.

Remark: Taking $s = 0$ in (NAG-SQC-ODE), we can automatically deduce the low resolution ODE associated with (NAG-SQC-DISCRETE).

4.2 The continuous/discrete rupture

As showed in the previous section, one can achieve accelerated convergence with the solution of the high resolution ODE (NAG-SQC-ODE) associated to Algorithm 1, without the need to add restriction on curvature.
As the Algorithm and the discrete Lyapunov function we use are discretization of continuous counterparts, one expect these to be similar. More precisely, ordering term by their dependence on s (1, $\mathcal{O}(\sqrt{s})$, $\mathcal{O}(s), \dots$), one expects the "main terms", namely the one with lower dependence on s, to appear both in continuous and discrete setting, while eventually, the discretization process will make appear new terms with higher dependency on s. In this section, we see why it is not necessarily the case when non convexity steps in.

Continuous/discrete Lyapunov comparison Recall the continuous Lyapunov used to prove Proposition 5.

$$\mathcal{E}(t) = \delta(F(X(t)) - F^*) + \frac{1}{2}\|\upsilon\dot{X}(t) + \lambda(X(t) - x^*) + \sqrt{s}\nabla F(X(t))\|^2 \tag{42}$$

where, depending on the values of γ, μ, L, we have $\upsilon \in [1, \frac{3}{2}], \delta \in [1, 3]$ and $\lambda \in [\sqrt{\mu}, \frac{9}{8}\sqrt{\mu}]$ (see the exact values in Appendix D). The discrete Lyapunov used to prove Theorem 2 can be written the following way:

$$E_n = F(x_n) - F^* + \frac{1}{2}\|\frac{(y_n - y_{n-1})}{\sqrt{s}} + \sqrt{\mu}(y_n - x^*) + \sqrt{s}\nabla F(y_{n-1})\|^2. \tag{43}$$

The fact that $\{E_n\}_{n\in\mathbb{N}}$ is a discretization of $\mathcal{E}$ appears here clearly.

Derivation difference In the continuous case, we studied $\dot{\mathcal{E}}$, where we aimed to show $\dot{\mathcal{E}}(t) \leqslant -\gamma\frac{\sqrt{\mu}}{2}\mathcal{E}(t)$. We emphasize the two following facts:

- Derivation of $\delta(F(X(t)) - F^*)$ leads to $\delta\langle\nabla F(X), \dot{X}\rangle$.
- Derivation of the norm term leads, among other terms, to $-\langle\nabla F(X), \dot{X}\rangle$ (up to a parameter).

Appropriate parameter tuning allows to cancel this terms, as can be seen in the proof of Proposition 5. In the discrete case, we study a discrete derivation $E_{n+1} - E_n$. Importantly, the two aforementioned derivation, that were the same in the continuous setting, will be different here.

- Discrete derivation of $F(x_{n+1}) - F^*$ leads to $F(x_{n+1}) - F(x_n)$.
- Discrete derivation of the norm term brings, among other terms, to $-\langle y_n - x_n, \nabla F(y_n)\rangle$.

More precisely, with an other view of the proof of Theorem 2 (see details in Appendix D.1), we get:

$$E_{n+1} - E_n \leqslant -\gamma\sqrt{\mu s}E_n + (1 - \gamma\sqrt{\mu s})\underbrace{(F(y_n) - F(x_n) + \langle\nabla F(y_n), x_n - y_n\rangle)}_{\text{Derivation difference}}$$
$$- \gamma(1 - \gamma\sqrt{\mu s})\sqrt{\frac{\mu}{s}}\|y_n - x_n\|^2. \tag{44}$$

Recall we want to end up with $E_{n+1} - E_n \leqslant -\gamma\sqrt{\mu s}E_n$. Then in (44), apart from the negative kinetic term, that can be matched with a continuous counterpart ($-\gamma\frac{\sqrt{\mu}v^2}{2}\|\dot{X}(t)\|^2$ term in line (202)), what remains is the difference of the two aforementioned derivation, i.e.:

$$F(y_n) - F(x_n) + \langle \nabla F(y_n), x_n - y_n \rangle. \tag{45}$$

In contrast to the continuous case, we can not just use parameter tuning to control this term. Convexity is exactly the assumption we need to get rid of this derivation difference term. However, if F is (a, L)-curvatured with $a < 0$, the best control we have is the following:

$$F(y_n) - F(x_n) + \langle \nabla F(y_n), x_n - y_n \rangle \leqslant -\frac{a}{2}\|x_n - y_n\|^2. \tag{46}$$

The lower below zero is a, or with other words the more non convexity we allow, the more this extra term can be big, up to a $+\frac{L}{2}\|x_n - y_n\|^2$ term in the L-smooth case ($a = -L$). If this term was appearing in the continuous case, the same restriction on the curvature as for Theorem 2 would be necessary. This highlights that when considering properties satisfied by the solution of a continuous system, the transfer of this properties to the discrete case via discretization can be hurt by non convexity.

Parallel with [13] This discretization problem occurring for non convex functions can be circumvented with the stochastic approach introduced in [13], later applied for (strongly) quasar convex functions minimization in [39]. Indeed, the iterations of their algorithm are defined as evaluations of a continuous process at some random times $T_0, T_1, \dots$ Then, a convergence result associated to this continuous process almost automatically transfer to the thus defined algorithm. Finally, this algorithm is practically usable as they show it can be written as a recursive formula of the form of Algorithm 1, with stochastic coefficients. Importantly with this view, the algorithm is not seen as a discretization of a continuous system in the sense that the algorithm trajectory would converge to the continuous one when a stepsize goes to zero. This explains that the non convexity negative impact described in this section does not occur.

5 Geometrical considerations

In this section, we investigate some new properties of strongly quasar convex functions. Moreover in the last subsection, we create a connection with the class of smooth PL functions (Definition 4). Firstly, we see how we can characterize strong quasar convexity in a similar but weaker way than strong convexity.

How weaker than strong convexity is strong quasar convexity ? Compared with strong convexity, the main difference is that we lose lots of *local information*. While for C^2 μ-strongly convex functions we have $\langle \nabla^2 F(x)y, y\rangle \geqslant \mu\|y\|^2$ for all $x, y \in \mathbb{R}^d$ (or equivalently, all eigenvalues of the Hessian matrix are above μ), we lose this regularity with strongly quasar convex functions. Actually, we see that we only have, on average, a similar regularity on the segment joining points $x \in \mathbb{R}^d$ and the minimizer x^*.

Proposition 6. *Let F be C^2 and (γ, μ)-strongly quasar convex for some $\gamma \in (0, 1]$, $\mu > 0$. Let $x \neq x^*$ and $t > 0$. Then:*

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 F(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds > \gamma\frac{\mu}{2}. \tag{47}$$

See appendix A.1 for the proof.

5.1 Dismissing local convexity argument for non C^2 strongly quasar convex functions

One may think that the property of uniqueness of the minimizer of strongly quasar convex functions induces that when we are close enough to the minimizer, the function is cuve-shaped and we avoid negative curvature, *i.e.* the function is locally convex around the minimizer. This is true if the function is C^2. The two following results holds under the assumption of a quadratic growth function with a unique minimizer, which is a weaker statement than strong quasar convexity (see Section 2.2).

Proposition 7. *Let F be C^2, with a unique minimizer x^*, and a μ-quadratic growth. Then, there exists $\eta > 0$ such that for all $x \in B(x^*, \eta)$, F is strongly convex.*

Sketch of Proof. The quadratic growth property around x^* implies that $\nabla^2 F(x^*)$ is definite positive. By continuity of the Hessian we get the result. Rigorous proof is deferred in appendix A.3. □

If we add a Hessian Lipschitz property, we can uniformly bound the distance from the minimizer that guaranties we do not reach a certain negative curvature.

Proposition 8. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be C^2, with a unique minimizer x^*, with μ-quadratic growth, and with its Hessian being ρ-Lipschitz. If for some $s \in \mathbb{R}$, we have $\|x - x^*\| \leq \frac{\mu - s}{\rho}$, then:*

$$\frac{(x - x^*)^T \nabla^2 F(x)(x - x^*)}{\|x - x^*\|^2} \geqslant s \tag{48}$$

Equivalently, $\|x - x^*\| \leq \frac{\mu - s}{\rho}$ implies that all eigenvalues of $\nabla^2 F(x)$ are above s.

Proof is in Appendix A.3.

Non C^2 functions When dropping the C^2 assumption, one can not ensure the local convexity property anymore, even when considering segments joining the minimizer (on which we have the stronger structure assumption). This is stated in the following result.

Proposition 9. *One can construct $f : [0, 1] \to \mathbb{R}$ strongly quasar convex with minimizer x^*, L-smooth, such that for all $x \neq x^*$, there exists $x_0 \neq x^*$ such that:*

$$|x^* - x_0| \leq |x^* - x| \text{ and } f''(x_0) = -L \tag{49}$$

Sketch of Proof. The idea behind the construction is the following: the function f is build such that its curvature alternates L and $-L$, given that the L curvature will occur more often than $-L$ to ensure that the function grows enough to be strongly quasar convex.
To create such a function, we define the following set

$$E := \bigcup_{n \geqslant 1} \left(\underbrace{\left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right)}_{\text{partition of [0,1]}} \cap \underbrace{\left[\frac{1}{2^n} + \frac{3}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right]}_{\text{subpart of partition}} \right) = \bigcup_{n \geqslant 1} \left[\frac{7}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right). \tag{50}$$

Then we define f on $[0, 1]$, such that:

$$f''(x) = \begin{cases} -L & \text{if } x \in E \\ L & \text{else.} \end{cases}$$

with $f'(0) = f(0) = 0$. f is clearly such that it will reach its lower curvature $-L$ in all vicinity of its minimizer. It remains to show that it is strongly quasar convex, which is done in Appendix A.4.

□

Other pathological examples The construction used to show Proposition 9 can be adapted to make other pathological behaviours. For example, consider μ-quadratic growth functions, *i.e.* which satisfy for all $x \in \mathbb{R}^d$, $f(x) - f^* \geqslant \frac{\mu}{2}\|x - x^*\|^2$. This defines functions lower bounded by a quadratic, but it may have critical points that are not (global) minima. However due to the local regularity around a global minimizer x^* offered by the quadratic growth, one may ask the following question:

Can we ensure that a smooth function with unique minimizer x^* satisfying μ-quadratic growth has no other critical points when sufficiently close to x^* ?

We address this question in appendix A.4, answering negatively by constructing a counter example based on the aforementioned construction.

5.2 A necessary condition for acceleration ? A link with Polyak Lojasiewicz condition

Recall that a function is μ-Polyak Lojasiewicz (μ-PL) if there exists $\mu > 0$ such that for all $x \in \mathbb{R}^d$, we have:

$$\frac{1}{2\mu}\|\nabla F(x)\|^2 \geqslant F(x) - F^*. \tag{51}$$

As already mentioned in section 2.1, according to [40], for μ-PL and L-smooth functions we can not get the acceleration phenomenon we witness for strongly convex functions. As we know that this acceleration can occur for the class of smooth strongly quasar convex function, it is interesting for a comprehension purpose to understand what is the gap between these functions and smooth PL functions. In other words:

What is missing for smooth PL functions to obtain acceleration ?

We propose an answer in this section. Strong quasar convexity implies uniqueness of minimizer, so for the sake of comparison we will consider the class of smooth PL function with a unique minimizer. Note that the function built in [40] to get the lower bound on smooth PL functions has a unique minimizer. There is thus still a point for comparison with this restricted class of smooth PL functions with unique minimizers, as gradient descent remains optimal when restricting to this class.

Theorem 4. *Suppose F is a μ-PL, L-smooth function with a unique minimizer for some $0 < \mu \leqslant L$. There exists $(\gamma, \mu') \in (0, 1] \times \mathbb{R}_+^*$ such that F is (γ,μ')-strongly quasar convex if and only if there exists some $a > 0$ such that F is satisfying the following* ***uniform acute angle condition****:*

$$\forall x \in \mathbb{R}^d, \quad 1 \geqslant \frac{\langle \nabla F(x), x - x^* \rangle}{\|\nabla F(x)\|\|x - x^*\|} \geqslant a > 0. \tag{UAAC}$$

The proof of Theorem 4 and complements are deferred in Appendix B, in which we give explicit parameters (γ, μ') depending on a, μ and L.
The (UAAC) condition can be interpreted in the following way: for all $x \in \mathbb{R}^d$, the descent direction $(-\nabla F(x))$ forms an acute angle with vector starting from x to the minimizer x^*. When it does not hold, following descent direction bring us to an orthogonal or opposite direction to the one that would make us closer to the minimizer.

Comments on Theorem 4 The need of (UAAC) condition in order to get acceleration is rather intuitive. Indeed it states that the momentum we accumulate is coherent, as it is directed toward the minimizer. For μ-PL and L-smooth functions, that are not necessarily satisfying the (UAAC) condition, we know that momentum does not allow to achieve accelerated rate [40]. Worse, in this case momentum appears to hurt the convergence rate. While the Polyak's Heavy Ball algorithm also leads to a linear convergence that is not better than gradient descent [12], it also deteriorate as we increase momentum. Thus, the fact that (UAAC) does not hold can make momentum hurt the convergence speed.

6 Numerical experiments details

6.1 Explicit expression of the function displayed in Figure 1

The function displayed in Figure 1 is $h : \mathbb{R}^2 \to \mathbb{R}$, and is built in the following way:

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right) \tag{52}$$

where $f(t) = t^2$ and

$$g(x_1, x_2) = \frac{1}{4N}\sum_{i=1}^{N}\left(a_i \sin(b_i x_1)^2 + c_i \cos(d_i x_2)^2\right) + 1 \tag{53}$$

with $N = 10$ and the $\{a_i\}_i$, $\{c_i\}_i$ are independently and uniformly distributed on $[0, 20]$, and the $\{b_i\}_i$, $\{d_i\}_i$ are independently and uniformly distributed on $[-25, 25]$.

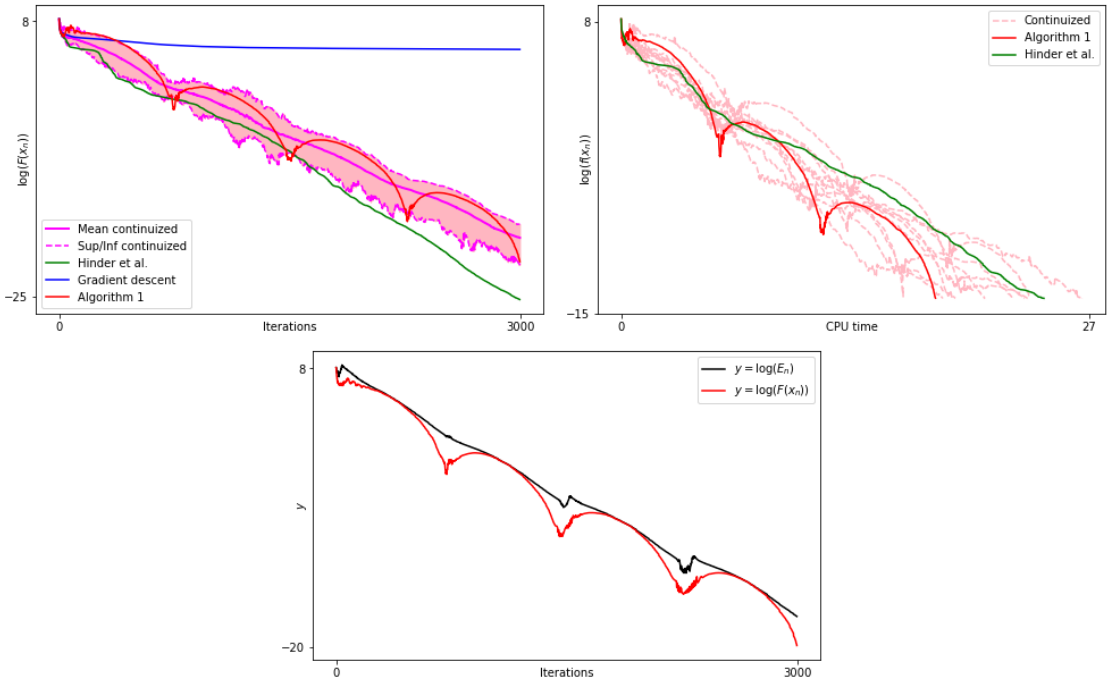

Figure 2: We compare the performance of an Algorithm using a line search procedure (Hinder et al. [21]), a stochastic algorithm (continuized [13]), and Algorithm 1. It is done iteration wise on the top left plot, while the top right compare the time needed to achieve a ε-solution. On the lowest plot we show the behaviour of Algorithm 1 with our choice of parameter in the presence of strong negative curvature regions.

6.2 Algorithm performance

We tested our algorithm on a function $h : \mathbb{R}^d \to \mathbb{R}$, $d = 100$, where

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right) \tag{54}$$

with $f(t) = t^2$, and

$$g(x_1, \dots, x_{100}) = \sum_{i=1}^{100} \left(a_i \sin(b_i x_i)^2\right) + 1 \tag{55}$$

where $\{a_i\}_i$ are independently and uniformly distributed on $[0, 1]$ and the $\{b_i\}_i$ are independently and uniformly distributed on $[-2.5, 2.5]$. This function is $(1, 2)$-strongly quasar convex, as f is $(1, 2)$-strongly quasar convex (see Proposition 10). Recall that this type of functions exhibits, by construction, lots of negative curvature. Hence, importantly, on these functions the assumption on the curvature of Theorem 2 is not satisfied.

Description of the experiments We performed numerical experiments on this function. We computed L at each iteration with the same backtracking process as in [21]. We compared the performance of Algorithm 1 using our choice of parameters with two other methods:

1. Algorithm 1 with line search computing the $(\alpha_n)_n$ [21].
2. Algorithm 1 with stochastic coefficients obtained using continuized framework [13, 39].

As the line search procedure induces more computational complexity, we compared the performance in two different ways: firstly iteration wise, and secondly we compared the CPU time needed to achieve an ε-precision, *i.e.* a point $\hat{x}$ such that $h(\hat{x}) - h^* < \varepsilon$, $\varepsilon > 0$.
The continuized framework leads to stochastic algorithm and thus to result of convergence are of

stochastic nature. Hence, to make the comparison with deterministic algorithms more relevant, we ran several times the algorithm. We make the following precision:

- For the iteration wise comparison, displayed on top left of figure 2, we ran 50 times the algorithm. We plot the mean trajectory (Mean continuized), as well as the infimum and supremum of all the trajectories along the iterations (Sup/Inf continuized). The pink zone between these two plots thus contains all the trajectories.
- For the CPU time needed to attain a ε-precision, displayed on top right of figure 2, we ran 10 times the algorithm. Here we simply plot 10 trajectories, corresponding each to a different run of the algorithm. We set $\varepsilon = 10^{-6}$.

A note on the using of Backtracking When using this backtracking, there is necessarily a divergence with our theoretical background. The L_n is computed such that $F(y_n - \frac{1}{L_n}\nabla F(y_n)) \leqslant F(y_n) - \frac{1}{2L_n}\|\nabla F(y_n))\|^2$. However to compute y_n, we need α_n, which needs the L_n we are willing to compute. We thus chose to compute α_n with the previous L_{n-1}.

Observations This precision being made, we can now state our observations.

1. The two top plots displayed in figure 2 are not very surprising: iteration wise, the binary search procedure offers a better speed. However when considering the CPU time needed to achieve a ε-precision, doing without line search allows for better performance as the iterations are less computationally heavy.
2. We observe empirically that a high amount of strong negative curvature encountered during the running of Algorithm 1 correlate with "bad behaviour" of the algorithm. The lowest display of Figure 2 is characteristic of the non monotone decreasing behaviour of the Lyapunov function E_n we can witness when the algorithm crosses strong negative curvature regions.

7 Conclusion

In this paper, we highlight that the Nesterov accelerated gradient algorithm may need curvature assumption to get accelerated rate in a non convex setting. As observed in previous works, we saw here that too strong negative curvature is difficult for this algorithm, at least regarding the nature of convergence results we are seeking. Interestingly, as it is the case for Polyak's Heavy Ball [35] in the strongly convex case, we saw that proving accelerated convergence of high resolution ODE associated to NAG in (strongly) quasar convex does not ensure convergence of discrete counterpart. Finally, it is still an open question whether there exists a deterministic algorithm achieving an accelerated rate on the class of smooth (strongly) quasar convex functions, without adding assumption, and without a subroutine to compute a parameter (as binary line search).

Acknowledgements

This work was supported by PEPR PDE-AI, the ANR MICROBLIND (grant ANR-21-CE48-0008) and the ANR Masdol (grant ANR-PRC-CE23).

References

[1] Vassilis Apidopoulos, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 187(1):151–193, 2021.

[2] Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, pages 1–43, 2022.

[3] Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, and Aude Rondepierre. Heavy Ball momentum for non-strongly convex optimization. *arXiv preprint arXiv:2403.06930*, 2024.

[4] Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.

[5] Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of the Heavy-Ball method under the łojasiewicz property. *Mathematical Programming*, 198(1):195–254, 2023.

[6] Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. On global convergence of ResNets: From finite to infinite width using linear parameterization. *Advances in Neural Information Processing Systems*, 35:16385–16397, 2022.

[7] Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[8] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[9] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pages 654–663. PMLR, 2017.

[10] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

[11] Patrick Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212, 2011.

[12] Marina Danilova, Anastasiia Kulakova, and Boris Polyak. Non-monotone behavior of the Heavy Ball method. In *Difference Equations and Discrete Dynamical Systems with Applications: 24th ICDEA, Dresden, Germany, May 21–25, 2018 24*, pages 213–230. Springer, 2020.

[13] Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendrikx, Pierre Gaillard, Laurent Massoulié, and Adrien Taylor. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. *Advances in Neural Information Processing Systems*, 34:28054–28066, 2021.

[14] Qiang Fu, Dongchu Xu, and Ashia Camage Wilson. Accelerated stochastic optimization methods under quasar-convexity. In *International Conference on Machine Learning*, pages 10431–10460. PMLR, 2023.

[15] Guillaume Garrigos. Square distance functions are Polyak-Łojasiewicz and vice-versa. *arXiv preprint arXiv:2301.10332*, 2023.

[16] Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

[17] Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2021.

[18] Charles Guille-Escuret, Adam Ibrahim, Baptiste Goujaud, and Ioannis Mitliagkas. Gradient descent is optimal under lower restricted secant inequality and upper error bound. *Advances in Neural Information Processing Systems*, 35:24893–24904, 2022.

[19] Sergey Guminov, Alexander Gasnikov, and Ilya Kuruzov. Accelerated methods for weakly-quasi-convex optimization problems. *Computational Management Science*, 20(1):36, 2023.

[20] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

[21] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pages 1894–1938. PMLR, 2020.

[22] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

[23] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

[24] Ilya A Kuruzov and Fedor S Stonyakin. Sequential subspace optimization for quasar-convex optimization problems with inexact gradient. In *International Conference on Optimization and Applications*, pages 19–33. Springer, 2021.

[25] Jasper CH Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614. IEEE, 2016.

[26] Jongmin Lee, Chanwoo Park, and Ernest Ryu. A geometric structure of acceleration and its role in making gradients small fast. *Advances in Neural Information Processing Systems*, 34:11999–12012, 2021.

[27] Bowen Li, Bin Shi, and Ya-xiang Yuan. Linear convergence of Nesterov-1983 with the strong convexity. *arXiv preprint arXiv:2306.09694*, 2023.

[28] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[29] 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.

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

[31] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[32] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[33] Yurii Nesterov. *Lectures on Convex Optimization*. 2018.

[34] Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4):773–810, 2021.

[35] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[36] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

[37] Jonathan W Siegel. Accelerated first-order methods: Differential equations and lyapunov functions. *arXiv preprint arXiv:1903.05671*, 2019.

[38] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[39] Jun-Kun Wang and Andre Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[40] Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing Polyak-Łojasiewicz functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2948–2968. PMLR, 2023.

[41] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

[42] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. *Advances in neural information processing systems*, 31, 2018.

[43] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *International Conference on Learning Representations*, 2019.

A Some properties on strongly quasar convex functions

In this section, we prove our claims about properties of strong quasar convex functions.

A.1 Strong convexity on average on segments joining the minimizer

Proposition. *Let F be (γ, μ)-strongly quasar convex. Let $x \neq x^*$ and $t > 0$. Then we have*

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 F(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geqslant \gamma\frac{\mu}{2} \tag{56}$$

Proof. Let $F : \mathbb{R}^d \to \mathbb{R}$ (γ, μ)-strongly quasar convex. Define, for $t \in \mathbb{R}_+$, the function $g(t) = F(x^* + t(x - x^*))$. We have $g'(t) = \langle \nabla F(x^* + t(x - x^*)), x - x^*\rangle$. By strong quasar convexity of F, we have

$$F(x^* + t(x - x^*)) + \frac{1}{\gamma}\langle \nabla F(x^* + t(x - x^*), x^* - (x^* + t(x - x^*))\rangle + \frac{\mu}{2}\|x^* - (x^* + t(x - x^*)\|^2 \leqslant F^* \tag{57}$$

$$\Rightarrow g(t) - \frac{t}{\gamma}g'(t) + \frac{\mu t^2}{2}\|x - x^*\|^2 \leqslant g(0) \tag{58}$$

$$\Rightarrow \gamma\frac{g(t) - g(0)}{t} + \frac{\gamma\mu t}{2}\|x - x^*\|^2 \leqslant g'(t) = \int_0^t g''(s)ds = \int_0^t \langle \nabla^2 F(x^* + s(x - x^*))x^* - x, x^* - x\rangle ds \tag{59}$$

Where for the last line we suppose $t > 0$. From this we deduce:

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 F(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geqslant \gamma\frac{\mu}{2} \tag{60}$$

□

In particular, the above reasoning remains true for non strongly quasar convex functions taking $\mu = 0$, inducing a on average convexity on segments joining minimizers:

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 F(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geq 0 \tag{61}$$

A.2 Synthetic strongly quasar convex example proof

In this section, we will use the useful following characterization of strong quasar convexity.

Lemma 2 ([21], lemma 11)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be differentiable function with a minimizer x^*, where the domain $\mathcal{X} \subset \mathbb{R}^d$ is open and convex. Then, the following two statements:*

$$f(tx^* + (1 - t)x) + t\left(1 - \frac{t}{2 - \gamma}\right)\frac{\gamma\mu}{2}\|x^* - x\|^2 \leqslant \gamma t f(x^*) + (1 - \gamma t)f(x), \forall x \in \mathcal{X}, t \in [0, 1] \tag{62}$$

$$f(x^*) \geqslant f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2, \forall x \in \mathcal{X} \tag{63}$$

are equivalent for all $\mu \geqslant 0$, $\gamma \in]0, 1]$.

Now let $f : \mathbb{R} \to \mathbb{R}$ such that it is (γ, μ)-strongly quasar convex and $f(0) = 0 = f^*$. Let $g : S^{d-1} \to \mathbb{R}$ differentiable and such that $g(x) \geqslant 1$ for all $x \in S^{d-1}$. We define

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right) \tag{64}$$

We have $h(x) \to 0$ as $x \to 0$, we extend h to 0 by continuity defining $h(0) = 0$. We clearly have $h^* = h(0) = 0$. We prove here the following statement.

Proposition 10. *h is (γ, μ)-strongly quasar convex.*

Proof. We use lemma 2 characterization of quasar strong convexity. We thus aim to show that for all $x \in \mathbb{R}^d$ and all $t \in [0, 1]$, we have

$$h(tx^* + (1-t)x) + t\left(1 - \frac{t}{2-\gamma}\right)\frac{\gamma\mu}{2}\|x^* - x\|^2 \leqslant \gamma t h(x^*) + (1-\gamma t)h(x) \tag{65}$$

First since $x^* = 0$ we have

$$h(tx^* + (1-t)x) = h((1-t)x) = f((1-t)\|x\|)\ g\left(\frac{x}{\|x\|}\right) \tag{66}$$

By strongly quasar convexity of f, we get

$$h((1-t)x) \leqslant \left((1-\gamma t)f(\|x\|) - t\left(1 - \frac{t}{2\gamma}\right)\frac{\mu\gamma}{2}\|x^* - x\|^2\right) g\left(\frac{x}{\|x\|}\right) \tag{67}$$

$$= (1-\gamma t)\underbrace{f(\|x\|)g\left(\frac{x}{\|x\|}\right)}_{=h(x)} - t\left(1 - \frac{t}{2\gamma}\right)\frac{\mu\gamma}{2}\|x^* - x\|^2 g\left(\frac{x}{\|x\|}\right) \tag{68}$$

We conclude by computing:

$$h((1-t)x) + t\left(1 - \frac{t}{2\gamma}\right)\frac{\mu\gamma}{2}\|x^* - x\|^2 \leqslant (1-\gamma t)h(x) + t\left(1 - \frac{t}{2\gamma}\right)\frac{\mu\gamma}{2}\|x^* - x\|^2\left(1 - g\left(\frac{x}{\|x\|}\right)\right) \tag{69}$$

Recalling $h^* = h(0) = 0$, we conclude using our condition $g \geq 1$, and we do have the characterization of (γ, μ)-strong quasar convexity. □

A.3 Local strong convexity around the minimizer for C^2 functions

Proposition. *Let F be C^2, with a unique minimizer x^*, with μ-quadratic growth. There exists $\eta > 0$ such that for all $x \in B(x^*, \eta)$, F is strongly convex.*

Proof.

Step 1: $\nabla^2 F(x^*)$ is definite positive We start by showing that $\nabla^2 F(x^*)$ is definite positive, *i.e.* for all $x \in \mathbb{R}^d \backslash \{0\}$ we have $\langle \nabla^2 F(x^*)x, x\rangle > 0$. Let $g(h) = F(x^* + h(x - x^*))$, $h \geq 0$ and $x \neq x^*$. We perform an order 2 Taylor development at 0 of g:

$$g(h) = g(0) + hg'(0) + \frac{h^2}{2}g''(0) + o(h^2) \tag{70}$$

$$\Leftrightarrow F(x^* + h(x - x^*) = F^* + h\underbrace{\langle \nabla F(x^*), x - x^*\rangle}_{=0} + \frac{h^2}{2}\langle \nabla^2 F(x^*)(x - x^*), x - x^*\rangle + o(h^2) \tag{71}$$

$$\Leftrightarrow F(x^* + h(x - x^*) - F^* = \frac{h^2}{2}\langle \nabla^2 F(x^*)(x - x^*), x - x^*\rangle + o(h^2) \tag{72}$$

As F is with μ-quadratic growth, we have:

$$F(x) - F^* \geqslant \frac{\mu}{2}\|x - x^*\|^2 \tag{73}$$

We thus have

$$\frac{h^2}{2}\langle \nabla^2 F(x^*)(x - x^*), x - x^*\rangle + o(h^2) \geqslant \frac{\mu h^2}{2}\|x - x^*\|^2 \tag{74}$$

$$\Rightarrow \frac{\langle \nabla^2 F(x^*)(x - x^*), x - x^*\rangle}{\|x - x^*\|^2} + \frac{2}{\|x - x^*\|^2}\frac{o(h^2)}{h^2} \geqslant \mu \tag{75}$$

Taking $h \to 0$, we get

$$\frac{\langle \nabla^2 F(x^*)(x - x^*), x - x^*\rangle}{\|x - x^*\|^2} \geqslant \mu \tag{76}$$

Taking $x = y + x^*$, we get that for all $y \in \mathbb{R}^d \backslash \{0\}$, we have $\langle \nabla^2 F(x^*)y, y\rangle \geqslant \mu\|y\|^2 > 0$. We showed that $\nabla^2 F(x^*)$ is definite positive.

Step 2: extension in a local vicinity We showed in step 1 that all eigenvalues of the Hessian matrix evaluated at x^* are strictly positive. As we assumed $\nabla^2 F$ is continuous, for all ε such that $0 < \varepsilon < \mu$, there exists $\eta > 0$ such that for all $x \in B(x^*, \eta)$, the eigenvalues of $\nabla^2 F(x)$ are above $\mu - \varepsilon$. This means that on this ball F is strongly convex, thus showing the claim. □

Proposition. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be C^2, with a unique minimizer x^*, with μ-quadratic growth, and with its Hessian being ρ-Lipschitz. If for some $s \in \mathbb{R}$, we have $\|x - x^*\| \leq \frac{\mu - s}{\rho}$, then:*

$$\frac{(x - x^*)^T \nabla^2 F(x)(x - x^*)}{\|x - x^*\|^2} \geqslant s \tag{77}$$

Equivalently, $\|x - x^*\| \leq \frac{\mu - s}{\rho}$ implies that all eigenvalues of $\nabla^2 F(x)$ are above s.

Proof. If the Hessian is ρ-Lipschitz, we have by definition:

$$\left|\left|\left|\nabla^2 F(x^*) - \nabla^2 F(x)\right|\right|\right| \leqslant \rho \|x - x^*\| \tag{78}$$

This induces that $\nabla^2 F(x) \in \mathbb{B}_{\rho\|x - x^*\|}\left(\nabla^2 F(x^*)\right)$. Then there exists $M \in \mathbb{R}^{d \times d}$ such that

$$\nabla^2 F(x) = \nabla^2 F(x^*) + M, \quad |||M||| \leqslant \rho \|x - x^*\| \tag{79}$$

Note that M is symmetric because $\nabla^2 F(x)$ and $\nabla^2 F(x^*)$ are symmetric. By the choice of the norm, we have:

$$-\rho \|x - x^*\| \|y\|^2 \leqslant yTMy \leqslant \rho \|x - x^*\| \|y\|^2 \tag{80}$$

Thus,

$$(x - x^*)^T \nabla^2 F(x)(x - x^*) \geqslant (x - x^*)^T \nabla^2 F(x^*)(x - x^*) - \rho \|x - x^*\|^3 \tag{81}$$

Now, we have already showed (76) that under our assumptions over F, we have $\langle \nabla^2 F(x^*)(x - x^*), x - x^* \rangle \geqslant \mu \|x - x^*\|^2$, $\forall x \in \mathbb{R}^d \backslash \{x^*\}$. Using this in (81) leads to:

$$\frac{(x - x^*)^T \nabla^2 F(x)(x - x^*)}{\|x - x^*\|^2} \geqslant (\mu - \rho \|x - x^*\|) \tag{82}$$

Finally, to ensure $\frac{(x-x^*)^T \nabla^2 F(x)(x-x^*)}{\|x-x^*\|^2} \geqslant s \in \mathbb{R}^d$, it suffices to have

$$\|x - x^*\| \leq \frac{\mu - s}{\rho} \tag{83}$$

□

A.4 No local convexity for non C^2 functions: a non convex pathological constructions

We construct in this section a type of functions that exhibit pathological non convex behaviour around their minimizer. This functions are defined on $\mathbb{R}$, so this pathological behaviour does not need several dimensions to happen.

A.4.1 Proof of proposition 9

Proposition. *One can construct $f : [0, 1] \to \mathbb{R}$ strongly quasar convex with minimizer x^*, L-smooth, such that for all $x \neq x^*$, there exists $x_0 \neq x^*$ such that:*

$$|x^* - x_0| \leq |x^* - x| \text{ and } f''(x_0) = -L \tag{84}$$

Proof. Let

$$E := \bigcup_{n \geqslant 1} \left(\underbrace{\left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right[}_{\text{partition of [0,1]}} \cap \underbrace{\left[\frac{1}{2^n} + \frac{3}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right]}_{\text{subpart of partition}} \right) = \bigcup_{n \geqslant 1} \left[\frac{7}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right[\tag{85}$$

and

$$E_n := \bigcup_{k \geqslant n+1} \left[\frac{7}{4}\frac{1}{2^k}, \frac{1}{2^{k-1}} \right[\tag{86}$$

Let f be a function defined on $[0,1]$, such that:

$$f''(x) = \begin{cases} -L & \text{if } x \in E \\ L & \text{else.} \end{cases}$$

We suppose $f'(0) = f(0) = 0$. f is clearly such that it will reach its lower curvature $-L$ in all vicinity of its minimizer. We now want to show that f is strongly quasar convex.
Suppose x is such that $\exists k \geqslant 1$, $x = \frac{1}{2^k}$. Then,

$$f'(x) = f'(x) - f'(0) = \int_{[0,x]} f''(s)ds = \int_{E_n} f''(s)ds + \int_{[0,x]\setminus E_n} f''(s)ds \tag{87}$$

By definition of f'', this simply becomes

$$f'(x) = -L\lambda(E_n) + L\lambda([0,x] \setminus E_n) = L(x - 2\lambda(E_n)) \tag{88}$$

Where $\lambda(.)$ is the Lebesgue measure. But by construction $\lambda(E_n) = \frac{1}{4}x$, which means

$$f'(x) = \frac{L}{2}x \tag{89}$$

Now suppose $x = \frac{1}{2^k} + \varepsilon$, where $k \geqslant 1$ and $0 < \varepsilon \leq \frac{3}{4}\frac{1}{2^k}$. This time we get

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^k}]} f''(s)ds + \int_{[\frac{1}{2^k},x]} f''(s)ds = \frac{L}{2}\frac{1}{2^k} + L\varepsilon \geqslant \frac{L}{2}(\frac{1}{2^k} + \varepsilon) = \frac{L}{2}x \tag{90}$$

Finally, suppose $x = \frac{1}{2^n} + \frac{3}{4}\frac{1}{2^n} + \varepsilon$, where $n \geqslant 1$ and $0 < \varepsilon < \frac{1}{4}\frac{1}{2^n}$.

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^n}+\frac{3}{4}\frac{1}{2^n}]} f''(s)ds + \int_{[\frac{1}{2^n}+\frac{3}{4}\frac{1}{2^n},x]} f''(s)ds = \frac{L}{2}\frac{1}{2^n} + \frac{3L}{4}\frac{1}{2^n} - L\varepsilon \geqslant \frac{L}{2^n} \tag{91}$$

But here $x \leqslant \frac{1}{2^{n-1}}$, which gives

$$f'(x) \geqslant \frac{L}{2}\frac{1}{2^{n-1}} = \frac{L}{2}x \tag{92}$$

Finally, for all $x \in [0,1]$, we have $f'(x) \geqslant \frac{L}{2}x$. To prove that this function is strongly quasar convex, let's remark by definition of f'' that $f(x) \leqslant \frac{L}{2}x^2$, and then using $f'(x) \geqslant \frac{L}{2}x$, we have that for any $\mu > 0$:

$$f(x) - \frac{\mu + L}{L}f'(x)x + \frac{\mu}{2}x^2 \leqslant \frac{L}{2}x^2 - \frac{\mu + L}{2}x^2 + \frac{\mu}{2}x^2 = 0 = f^* \tag{93}$$

In words we showed that f is $(\frac{L}{\mu+L}, \mu)-$strongly quasar convex. In conclusion, we created a 1d function which is strongly quasar convex, and for which there exists no neighbourhood around minimizer such that negative curvature is excluded. □

A.4.2 Quadratic growth and local uniqueness of critical points ?

We recall that a function satisfies μ-quadratic growth hypothesis if it satisfies

$$f(x) - f^* \geqslant \frac{\mu}{2}\|x - x^*\|^2 \tag{94}$$

Suppose such a function has a unique minimizer. Regarding the local regularity around x^* provided by this hypothesis, it is not obvious whether there exists a vicinity around this minimizer such that there is no other critical points. We provide a counter example.

Proposition 11. *There exists a quadratic growth function $f : [0,1] \to \mathbb{R}$ such that in all vicinity of its minimizer, f has an infinite amount of critical points.*

Proof. Let

$$E := \bigcup_{n\geqslant 1} \left(\left[\frac{1}{2^n}, \frac{1}{2^{n-1}} \right[\cap \left[\frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n}, \frac{1}{2^{n-1}} \right] \right) = \bigcup_{n\geqslant 1} \left[\frac{3}{2}\frac{1}{2^n}, \frac{1}{2^{n-1}} \right[\tag{95}$$

and

$$E_n := \bigcup_{k\geqslant n+1} \left[\frac{3}{2}\frac{1}{2^k}, \frac{1}{2^{k-1}} \right[\tag{96}$$

Let f be a function defined on $[0,1]$, such that:

$$f''(x) = \begin{cases} -1 & \text{if } x \in E \\ 1 & \text{else.} \end{cases}$$

We suppose $f'(0) = f(0) = 0$. Suppose x is such that $\exists n \geqslant 0$, $x = \frac{1}{2^n}$. Then,

$$f'(x) = f'(x) - f'(0) = \int_{[0,x]} f''(s)ds = \int_{E_n} f''(s)ds + \int_{[0,x]\setminus E_n} f''(s)ds \tag{97}$$

By definition of f'', this simply becomes

$$f'(x) = -\lambda(E_n) - \lambda([0,x] \setminus E_n) = x - 2\lambda(E_n) \tag{98}$$

Where $\lambda(.)$ is the Lebesgue measure. But by construction $\lambda(E_n) = \frac{1}{2}x$, which means

$$f'(x) = 0 \tag{99}$$

This means our function has infinite amount of critical point in all vicinity of minimizer. We want now to show that f is a quadratic growth function.
We suppose $x = \frac{1}{2^k} + \varepsilon$, where $k \geqslant 1$ and $0 < \varepsilon \leq \frac{1}{2}\frac{1}{2^k}$. This time we get

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^k}]} f''(s)ds + \int_{[\frac{1}{2^k},x]} f''(s)ds = \varepsilon \tag{100}$$

Now suppose $x = \frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n} + \varepsilon$, where $n \geqslant 1$ and $0 < \varepsilon < \frac{1}{2}\frac{1}{2^n}$.

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n}]} f''(s)ds + \int_{[\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n},x]} f''(s)ds = \frac{1}{2}\frac{1}{2^n} - \varepsilon \tag{101}$$

We have then

$$f\left(\frac{1}{2^{n-1}}\right) - f\left(\frac{1}{2^n}\right) = 2\int_{[\frac{1}{2^n},\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n}]} f'(s)ds = 2\int_{[\frac{1}{2^n},\frac{3}{2}\frac{1}{2^n}]} sds = \frac{5}{4}\frac{1}{2^{2n}} \tag{102}$$

Summing over n we get

$$f\left(\frac{1}{2^{n-1}}\right) = \frac{5}{4}\sum_{i\geqslant n}\frac{1}{2^{2n}} = \frac{5}{4}\frac{\frac{1}{4^n}}{1-\frac{1}{4}} = \frac{5}{12}\frac{1}{2^{2n-2}} = \frac{5}{12}\left(\frac{1}{2^{n-1}}\right)^2 \tag{103}$$

If $x = \frac{1}{2^k} + \varepsilon$, where $k \geqslant 1$ and $0 < \varepsilon \leq \frac{1}{2}\frac{1}{2^k}$, We have

$$f(x) - f\left(\frac{1}{2^n}\right) = \int_{[\frac{1}{2^n},x]} f'(s)ds = \frac{1}{2}\left(x^2 - \frac{1}{2^{2n}}\right) \tag{104}$$

$$\Rightarrow f(x) = \frac{5}{12}\frac{1}{2^{2n}} + \frac{1}{2}x^2 - \frac{6}{12}\frac{1}{2^{2n}} = \frac{5}{12}x^2 + \frac{1}{12}x^2 - \frac{1}{12}\frac{1}{2^{2n}} \tag{105}$$

However as $x \geqslant \frac{1}{2^n}$, we get that $f(x) \geqslant \frac{5}{12}x^2$.
Now suppose $x = \frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n} + \varepsilon$, where $n \geqslant 1$ and $0 < \varepsilon < \frac{1}{2}\frac{1}{2^n}$. We show by contradiction that $f(x) \geqslant \frac{5}{12}x^2$. Suppose $f(x) < \frac{5}{12}x^2$. Note that as we showed

$$f'(x) = \frac{1}{2}\frac{1}{2^n} - \varepsilon \tag{106}$$

we have

$$\frac{5}{6}x - f'(x) = \frac{15}{12}\frac{1}{2^n} + \frac{5}{6}x_2 - \frac{1}{2}\frac{1}{2^n} + \varepsilon = \frac{6}{12}\frac{1}{2^n} + \frac{9}{12} \geqslant 0 \tag{107}$$

So we have both:

$$f(x) < \frac{5}{12}x^2 \text{ and } f'(x) \leqslant \left(\frac{5}{12}x^2\right)' = \frac{5}{6}x \tag{108}$$

The second inequality holds for all y in $[\frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n}, \frac{1}{2^{n-1}}]$, inducing that $f(y) < \frac{5}{12}y^2$. However, we already showed that $f(\frac{1}{2^{n-1}}) = \frac{5}{12}\left(\frac{1}{2^{n-1}}\right)^2$, leading to a contradiction. We conclude that $f(x) \geqslant \frac{5}{12}x^2$. We then showed that

$$f(x) = f(x) - f^* \geqslant \frac{\frac{5}{12}}{2}x^2 = \frac{\frac{5}{12}}{2}(x - x^*)^2 \tag{109}$$

or in words, f has infinite amount of critical point in each vicinity of minimizer and is $\frac{5}{12}$-quadratic growth.

□

B A PL-Strongly quasar convex link

In this section we characterize the difference between smooth PL functions (Definitions 2 and 3 respectively) and smooth, strongly quasar convex functions. The idea of the following discussion is to use relation of μ-PL functions with intermediate conditions that we can relate to strongly quasar convex functions.

Remark 1 We can not claim that all the following lemmas are new. We indicated a citation when we were aware that a result already exists in the literature.

Remark 2 We will introduce geometrical conditions that can hold considering projection onto the set of minimizers. As we aim to show a link with strong quasar convexity, we will restrict ourselves to functions with a unique minimizer. This means that some of the definitions we will introduce are specifically here restricted to this unique minimizer case.

For the first lemma, we introduce the following condition.

Definition 7 (Error Bound)**.** *$F : \mathbb{R}^d \mapsto \mathbb{R}$ is θ-Error Bound (θ-EB) if*

$$\forall x \in \mathbb{R}^d, \quad \|\nabla F(x)\| \geqslant \theta\|x - x^*\|$$

Lemma 3. *A μ-PL function is μ-EB. A θ-EB and L-smooth function is $\frac{\theta}{L}$-PL.*

It is shown in [23]. As it is short we will give the proof.

Proof.

PL $\Rightarrow$ EB Let F be μ-PL. By definition and using the fact that a μ-PL function is also μ-quadratic growth (see [15] theorem 11, or [23]):

$$\|\nabla F(x)\|^2\frac{1}{2\mu} \geqslant F(x) - F^* \geqslant \frac{\mu}{2}\|x - x^*\|^2 \Rightarrow \|\nabla F(x)\| \geqslant \mu\|x - x^*\| \tag{110}$$

EB + L-smooth $\Rightarrow$ PL Suppose F is μ-EB and L-smooth. We have

$$\|\nabla F(x)\| \overset{(EB)}{\geqslant} \theta\|x - x^*\| \overset{\text{(L-smooth)}}{\geqslant} \frac{2\theta}{L}(F(x) - F^*) \tag{111}$$

Where we use the fact that a L-smooth function verifies $F(x) - F^* \leqslant \frac{L}{2}\|x - x^*\|^2$. □

We now introduce the notion of RSI functions [41].

Definition 8. *$F : \mathbb{R}^d \mapsto \mathbb{R}$ is ν-RSI if*

$$\langle \nabla F(x), x - x^* \rangle \geqslant \nu\|x - x^*\|^2, \quad \forall x \in \mathbb{R}^d$$

Using Cauchy Schwartz, we immediately see that ν-RSI implies ν-EB. We show that up to the a supplementary condition, the converse also holds.

Lemma 4. *Suppose F is satisfying the following **uniform acute angle condition**:*

$$\forall x \in \mathbb{R}^d, \quad 1 \geqslant \frac{\langle \nabla F(x), x - x^* \rangle}{\|\nabla F(x)\|\|x - x^*\|} \geqslant a > 0 \tag{UAAC}$$

then

$$F \text{ is } \theta\text{-EB} \Rightarrow F \text{ is } \theta a\text{-RSI}$$

Proof. We have

$$\theta\|x - x^*\| \overset{\theta-(EB)}{\leq} \|\nabla F(x)\| \overset{(UAAC)}{\leq} \frac{\langle \nabla F(x), x - x^* \rangle}{a\|x - x^*\|} \Rightarrow \theta a\|x - x^*\|^2 \leqslant \langle \nabla F(x), x - x^* \rangle \tag{112}$$

□

In the following result, we establish a link with a last intermediate condition, namely verifying quadratic growth and (non strong) quasar convexity.

Lemma 5. *Let F be L-smooth. Then:*

1. *(F is (γ, μ)-strongly quasar convex) $\Rightarrow$ (F is $\frac{\gamma\mu}{2-\gamma}$-RSI)*
2. *(F is ν-RSI) + ($\gamma < \frac{2\nu}{L}$) $\Rightarrow$ (F is $\left(\gamma, \frac{\nu}{\gamma} - \frac{L}{2}\right)$-strongly quasar convex)*

Proof.

Point 1. Using definition of strong quasar convexity, and the fact that it implies $\frac{\mu\gamma}{2-\gamma}$ quadratic growth (Proposition 1, 2.), we have:

$$\langle \nabla F(x), x - x^* \rangle \geqslant \gamma(F - F^*) + \frac{\gamma\mu}{2}\|x - x^*\|^2 \geqslant \frac{\gamma^2\mu}{2(2-\gamma)}\|x - x^*\|^2 + \frac{\gamma\mu}{2}\|x - x^*\|^2 \tag{113}$$

$$= \frac{\gamma\mu}{2-\gamma}\|x - x^*\|^2 \tag{114}$$

Point 2. We start with the definition of ν-RSI:

$$\langle \nabla F(x), x - x^* \rangle \geqslant \nu\|x - x^*\|^2 \Rightarrow 0 \geqslant \frac{1}{\gamma}\langle \nabla F(x), x^* - x \rangle + \frac{\nu}{\gamma}\|x - x^*\|^2 \tag{115}$$

$$\Rightarrow -\frac{L}{2}\|x - x^*\|^2 \geqslant \frac{1}{\gamma}\langle \nabla F(x), x^* - x \rangle + \left(\frac{\nu}{\gamma} - \frac{L}{2}\right)\|x - x^*\|^2 \tag{116}$$

Where $\gamma \in (0, 1]$ is to be precised. L-smooth property implies $F(x) - F^* \leqslant \frac{L}{2}\|x - x^*\|^2$, thus we have:

$$F^* \geqslant F(x) + \frac{1}{\gamma}\langle \nabla F(x), x^* - x \rangle + \left(\frac{\nu}{\gamma} - \frac{L}{2}\right)\|x - x^*\|^2 \tag{117}$$

Hence, choosing γ such that $\frac{\nu}{\gamma} - \frac{L}{2} > 0 \Rightarrow \gamma < \frac{2\nu}{L}$, we have that F is $\left(\gamma, \frac{\nu}{\gamma} - \frac{L}{2}\right)$-strongly quasar convex. □

Note that we did not really need gradient Lipschitz property to hold, rather a weaker upper quadratic growth condition (this also holds for Lemma 3):

$$\frac{L_0}{2}\|x-\hat{x}\|^2 \geqslant F(x)-F^* \tag{118}$$

An important point here is that L_0 may be significantly lower that L. See [17] for a discussion about alternative upper conditions.

With all these lemmas, we are ready to state our result.

Theorem 5. *Let F be L-smooth and μ-PL with a unique minimizer. Then we have that there exists $\gamma, \mu' > 0$ such that F is (γ, μ')-strongly quasar convex if and only if for some $a \in (0,1]$, F satisfies the following* ***uniform acute angle condition****:*

$$\forall x \in \mathbb{R}^d, \quad 1 \geqslant \frac{\langle \nabla F(x), x-x^* \rangle}{\|\nabla F(x)\|\|x-x^*\|} \geqslant a > 0 \tag{UAAC}$$

In particular if (UAAC) holds, then as long as $\gamma < \frac{2\mu a}{L}$, F is $(\gamma, \frac{\mu a}{\gamma} - \frac{L}{2})$-strongly quasar convex.

Proof. Let F be L-smooth and μ-PL. We have by Lemma 3 that F is μ-EB. By Lemma 4 F is μa-RSI. By Lemma 5, F is $\left(\gamma, \frac{\mu a}{\gamma} - \frac{L}{2}\right)$-strongly quasar convex as long as $\gamma > \frac{2\mu a}{L}$. Obviously, if there exists $x \neq x^*$ such that (UAAC) does not hold, then RSI can not hold. As strong quasar convexity implies RSI (Lemma 5), this is a necessary condition for theorem 5 to hold. □

Finally, we show that strong quasar convexity implies the PL condition without the need of adding assumptions.

Proposition 12. *Let F be (γ, μ)-strongly quasar convex. It is then $\mu\gamma^2$-PL.*

Proof. We have

$$\frac{1}{2}\|\frac{1}{\gamma\sqrt{\mu}}\nabla F(x) + \sqrt{\mu}(x^*-x)\|^2 = \frac{1}{2\gamma^2\mu}\|\nabla F(x)\|^2 + \frac{1}{\gamma}\langle \nabla F(x), x^*-x\rangle + \frac{\mu}{2}\|x-x^*\|^2 \tag{119}$$

Writting the definition of (γ, μ)-strong quasar convexity, we have

$$F^* \geqslant F(x) + \frac{1}{\gamma}\langle \nabla F(x), x^*-x\rangle + \frac{\mu}{2}\|x-x^*\|^2 \tag{120}$$

$$= F(x) + \frac{1}{2}\|\frac{1}{\gamma\sqrt{\mu}}\nabla F(x) + \sqrt{\mu}(x^*-x)\|^2 - \frac{1}{2\gamma^2\mu}\|\nabla F(x)\|^2 \tag{121}$$

$$\Rightarrow \frac{1}{2\gamma^2\mu}\|\nabla F(x)\|^2 \geqslant F(x) - F^* \tag{122}$$

□

One can find a summary of the previous discussion in figure 3.

C Proofs of section 3

C.1 Differentiable strongly quasar convex

C.1.1 Proof of Theorem 2

In this section, we detail the proof of Theorem 2 whose statement is recalled here: let F be a (γ, μ)-strongly quasar convex function for some $(\gamma, \mu) \in (0,1] \times \mathbb{R}_+^*$ having a (ρ, L) curvature for some $L > 0$ and $\rho \leqslant L$. Let $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 1:

$$\begin{aligned} y_n &= \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} &= y_n - s\nabla F(y_n) \\ z_{n+1} &= \beta_n z_n + (1-\beta_n) y_n - \eta_n \nabla F(y_n) \end{aligned}$$

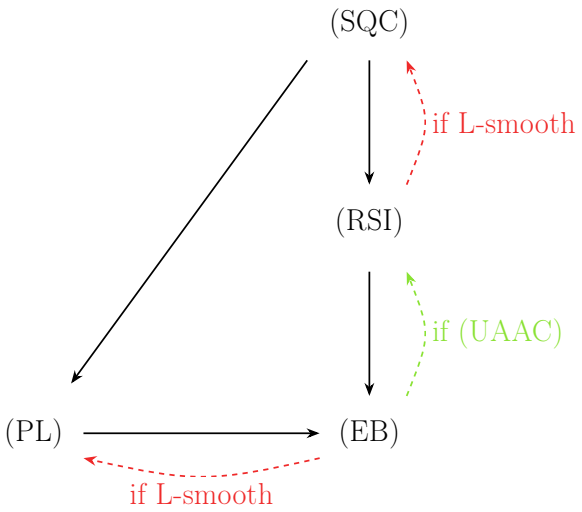

Figure 3: Summary of the Lemmas of Appendix B. See Definition 5 for SQC (strongly quasar convex), Definition 8 for RSI, Definition 7 for EB (error bound), and Definition 4 for PL (Polyak-Łojasiewcz). Solid lines are implications that hold without the need of adding another assumption. Red dashed lines are implications that hold under L-smooth assumption, while the green dashed line is for implication holding under the (UAAC) condition.

with parameters

$$s \leqslant \frac{1}{L}, \quad \alpha_n = \frac{1}{1+\sqrt{\mu s}} := \alpha, \quad \beta_n = 1 - \gamma\sqrt{\mu s} := \beta, \quad \eta_n = \frac{\sqrt{s}}{\sqrt{\mu}} := \eta.$$

Assuming that $\rho \geqslant -\gamma\sqrt{\frac{\mu}{s}}$ we want to prove that:

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant \frac{2}{\gamma}\left(1-\gamma\sqrt{\mu s}\right)^n \left(F(x_0) - F^*\right) \tag{123}$$

where $F^* = \min\ F$. Let x^* be the unique minimizer of F. We introduce the following Lyapunov energy:

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|z_n - x^*\|^2 \tag{124}$$

The main idea of the proof consists in finding parameters and conditions such that the following inequality holds

$$E_{n+1} - E_n \leqslant cE_n \tag{125}$$

with $c < 0$ being as small as possible. We will then deduce the convergence rate (123) by induction.

Step 1. Since:

$$E_{n+1} - E_n = F(x_{n+1}) - F(x_n) + \frac{\mu}{2}\|z_{n+1} - x^*\|^2 - \frac{\mu}{2}\|z_n - x^*\|^2, \tag{126}$$

let us start by considering the right term:

$$\begin{aligned}
\Delta_n &= \|z_{n+1} - x^*\|^2 - \|z_n - x^*\|^2 \\
&= \|\beta z_n + (1-\beta)y_n - \eta\nabla F(y_n) - x^*\|^2 - \|z_n - x^*\|^2 \\
&= (\beta^2 - 1)\|z_n - x^*\|^2 + (1-\beta)^2\|y_n - x^*\|^2 + \eta^2\|\nabla F(y_n)\|^2 \\
&\quad +2\beta\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla F(y_n)\rangle - 2(1-\beta)\eta\langle \nabla F(y_n), y_n - x^*\rangle
\end{aligned}$$

by construction of Algorithm 1. The tricky part here is to control the first scalar product: using the definition of Algorithm 1, we can rewrite $z_n = y_n + \frac{\alpha}{1-\alpha}(y_n - x_n)$, and thus

$$\begin{aligned}
&\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla F(y_n)\rangle \\
&= \langle y_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla F(y_n)\rangle + \frac{\alpha}{1-\alpha}\langle y_n - x_n, (1-\beta)(y_n - x^*) - \eta\nabla F(y_n)\rangle \\
&= (1-\beta)\|y_n - x^*\|^2 - \eta\langle y_n - x^*, \nabla F(y_n)\rangle - \frac{\alpha}{1-\alpha}\eta\langle y_n - x_n, \nabla F(y_n)\rangle \\
&+ \frac{\alpha}{1-\alpha}(1-\beta)\langle y_n - x_n, y_n - x^*\rangle
\end{aligned}$$

Observe now that applying the relation $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ to $a = y_n - x^*$ and $b = \frac{\alpha}{1-\alpha}(y_n - x_n)$, we get:

$$\frac{\alpha}{1-\alpha}\langle y_n - x_n, y_n - x^*\rangle = \frac{1}{2}\|z_n - x^*\|^2 - \frac{1}{2}\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{1}{2}\|y_n - x^*\|^2, \tag{127}$$

so that:

$$\begin{aligned}
&\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla F(y_n)\rangle \\
&= \frac{1-\beta}{2}\left(\|z_n - x^*\|^2 + \|y_n - x^*\|^2 - \left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2\right) - \eta\langle y_n - x^*, \nabla F(y_n)\rangle \\
&- \frac{\alpha}{1-\alpha}\eta\langle y_n - x_n, \nabla F(y_n)\rangle
\end{aligned}$$

and

$$\begin{aligned}
\Delta_n \;=\; & -(1-\beta)\|z_n - x^*\|^2 + (1-\beta)\|y_n - x^*\|^2 + \eta^2\|\nabla F(y_n)\|^2 \\
& -\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - 2\frac{\alpha\beta\eta}{1-\alpha}\langle\nabla F(y_n), y_n - x_n\rangle - 2\eta\langle\nabla F(y_n), y_n - x^*\rangle.
\end{aligned}$$

Reinjecting Δ_n in the expression of $E_{n+1} - E_n$ and by definition of the Lyapunov energy E_n, we then get:

$$\begin{aligned}
E_{n+1} - E_n = & -(1-\beta)E_n + F(x_{n+1}) - F^* - \beta\left(F(x_n) - F^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 + \frac{\mu}{2}\eta^2\|\nabla F(y_n)\|^2 \\
& - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle\nabla F(y_n), y_n - x_n\rangle - \mu\eta\langle\nabla F(y_n), y_n - x^*\rangle.
\end{aligned} \tag{128}$$

Step 2. Let us now prove that for any $n \in \mathbb{N}$, we have: $E_{n+1} - E_n \leqslant -(1-\beta)E_n$ for some well-chosen values of the parameters β, η and α.

Remember that F is assumed strongly quasar convex, hence:

$$\forall n \in \mathbb{N},\ \langle\nabla F(y_n), y_n - x^*\rangle \geqslant \gamma(F(y_n) - F^*) + \frac{\gamma\mu}{2}\|y_n - x^*\|^2 \tag{129}$$

and L-smooth which induces that:

$$\forall s \leqslant \frac{1}{L},\ \forall n \in \mathbb{N},\ \frac{s}{2}\|\nabla F(y_n)\|^2 \leqslant F(y_n) - F(x_{n+1}) \tag{130}$$

Reinjecting these two inequalities into $E_{n+1} - E_n$, we then get:

$$\begin{aligned}
E_{n+1} - E_n &\leqslant -(1-\beta)E_n + F(x_{n+1}) - F^* - \beta\left(F(x_n) - F^*\right) + \frac{\mu}{s}\eta^2(F(y_n) - F(x_{n+1})) \\
&\quad - \frac{\mu\beta(1-\beta)}{2}\left(\frac{\alpha}{1-\alpha}\right)^2 \|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle \nabla F(y_n), y_n - x_n\rangle - \eta\mu\gamma(F(y_n) - F^*) \\
&\quad + \frac{\mu}{2}(1-\beta-\gamma\eta\mu)\|y_n - x^*\|^2 \\
&\leqslant -(1-\beta)E_n + (\frac{\mu}{s}\eta^2 - \gamma\mu\eta)(F(y_n) - F^*) + (1 - \frac{\mu}{s}\eta^2)(F(x_{n+1}) - F^*) \\
&\quad - \beta(F(x_n) - F^*) - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle \nabla F(y_n), y_n - x_n\rangle - \frac{\mu\beta(1-\beta)}{2}\left(\frac{\alpha}{1-\alpha}\right)^2 \|y_n - x_n\|^2 \quad (131) \\
&\quad + \frac{\mu}{2}(1-\beta-\gamma\eta\mu)\|y_n - x^*\|^2
\end{aligned}$$

Choosing now $\eta = \sqrt{\frac{s}{\mu}}$ and $\beta = 1 - \gamma\eta\mu = 1 - \gamma\sqrt{\mu s}$ to cancel out the terms in $F(x_{n+1}) - F^*$ and $\|y_n - x^*\|^2$, we deduce:

$$\begin{aligned}
E_{n+1} - E_n &\leqslant -\gamma\sqrt{\mu s}E_n + (1-\gamma\sqrt{\mu s})\left(F(y_n) - F(x_n)\right) \\
&\quad + (1-\gamma\sqrt{\mu s})\sqrt{\mu s}\frac{\alpha}{1-\alpha}\left(\langle \nabla F(y_n), x_n - y_n\rangle - \frac{\alpha}{1-\alpha}\frac{\gamma\mu}{2}\|y_n - x_n\|^2\right). \quad (132)
\end{aligned}$$

Suppose additionally that the lower curvature is bounded from below by ρ, hence:

$$\forall n \in \mathbb{N},\ F(y_n) + \langle \nabla F(y_n), x_n - y_n\rangle + \frac{\rho}{2}\|x_n - y_n\|^2 \leqslant F(x_n),$$

or equivalently:

$$\forall n \in \mathbb{N},\ \langle \nabla F(y_n), x_n - y_n\rangle \leqslant F(x_n) - F(y_n) - \frac{\rho}{2}\|x_n - y_n\|^2.$$

Injecting the very last inequality into (132) we get:

$$\begin{aligned}
E_{n+1} - E_n &\leqslant -\gamma\sqrt{\mu s}E_n + (1-\gamma\sqrt{\mu s})\left(1 - \sqrt{\mu s}\frac{\alpha}{1-\alpha}\right)\left(F(y_n) - F(x_n)\right) \\
&\quad + (1-\gamma\sqrt{\mu s})\frac{\sqrt{\mu s}}{2}\frac{\alpha}{1-\alpha}\left(-\rho - \frac{\alpha}{1-\alpha}\gamma\mu\right)\|y_n - x_n\|^2
\end{aligned}$$

Lastly, choose $\alpha = \frac{1}{1+\sqrt{\mu s}}$ to cancel out the term in $F(y_n) - F(x_n)$, so that we finally get:

$$E_{n+1} - E_n = -\gamma\sqrt{\mu s}E_n + \frac{1-\gamma\sqrt{\mu s}}{2}\left(-\rho - \gamma\frac{\sqrt{\mu}}{\sqrt{s}}\right)\|y_n - x_n\|^2 \leqslant -\gamma\sqrt{\mu s}E_n$$

provided that the lower curvature satisfies: $\rho \geqslant -\gamma\frac{\sqrt{\mu}}{\sqrt{s}}$.

Step 3. To conclude, we proved that:

$$E_{n+1} - E_n \leqslant -\gamma\sqrt{\mu s}E_n \Rightarrow E_{n+1} \leqslant (1-\gamma\sqrt{\mu s})E_n \quad (133)$$

By induction:

$$\forall n \in \mathbb{N},\ E_{n+1} \leqslant (1-\gamma\sqrt{\mu s})^{n+1}\, E_0 = (1-\gamma\sqrt{\mu s})^{n+1}\left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) \quad (134)$$

In the last equality we use that by definition of Algorithm 1 $x_0 = z_0$. We then use the fact that (γ, μ) strong convexity implies $\frac{\gamma\mu}{2-\gamma}$-quadratic growth (corollary 1 [21]), *i.e.*

$$\forall x \in \mathbb{R}^d,\ F(x) - F^* \geqslant \frac{\gamma\mu}{2(2-\gamma)}\|x - x^*\|^2 \quad (135)$$

which finally yields to:

$$F(x_n) - F^* \leqslant \frac{2}{\gamma}\left(1-\gamma\sqrt{\mu s}\right)^n \left(F(x_0) - F^*\right). \quad (136)$$

C.1.2 2 points scheme version of algorithm 1

Here we build upon the work [26], where they show there exists an equivalence between a 3 points and 2 points scheme version of Nesterov Accelerated gradient. We want to deduce a 2 points scheme from our 3 points one, but we can not directly apply their result because the 3 points algorithm they consider is slightly different.

Proposition 13. *The algorithm*

$$\begin{cases} y_n = \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} = y_n - s\nabla F(y_n) \\ z_{n+1} = \beta_n z_n + (1-\beta_n) y_n - \eta_n \nabla F(y_n) \end{cases} \tag{137}$$

can be written as the following 2 points scheme

$$\begin{cases} y_n = x_n + \frac{1-\alpha_n}{1-\alpha_{n-1}}\alpha_{n-1}\beta_{n-1}(x_n - x_{n-1}) + (1-\alpha_n)\left(\frac{\eta_{n-1}}{s} - \frac{\alpha_{n-1}\beta_{n-1}}{1-\alpha_{n-1}} - 1\right)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla F(y_n) \end{cases} \tag{138}$$

Proof. We adapt the proof of Lemma 2 from [26], which is mainly based on Thales theorem. In their result, the α_n is $\frac{\varphi_n}{\varphi_{n-1}}$ and $\beta_n z_n + (1-\beta_n) y_n$ is z_n. The main idea is that if a vector u is colinear to v, *i.e.* $\exists \lambda \in \mathbb{R}\ u = \lambda v$, then $\lambda = \frac{\|u\|}{\|v\|}$, which we can find using Thales theorem. As in the original proof, let us rewrite (137) as:

$$\begin{cases} x_n = \alpha_n x_{n-1}^+ + (1-\alpha_n) z_n \\ z_{n+1} = \beta_n z_n + (1-\beta_n) x_n - \eta_n \nabla F(x_n) \end{cases} \tag{139}$$

where $x_{n-1}^+ = x_{n-1} - s\nabla F(x_{n-1})$. We suppose $\nabla F(x_{n-1}) \neq 0$ (non degenerate case). We set $v_n = \beta_n z_n + (1-\beta_n) x_n$, and let's set A on the $[x_{n-1}^+ x_n]$ segment such that $Ax_{n+1} \parallel x x_n^+$. Let B on $x_{n+1}^+ x_n^+ \cap v_n z_{n+1}$. Since $Ax_{n+1} \parallel Bz_{n+1}$, we have by Thales theorem:

$$\frac{\|B-A\|}{\|B-x_n^+\|} = \frac{\|z_{n+1} - x_{n+1}\|}{\|z_{n+1} - x_n^+\|} := (\star) \tag{140}$$

Then, by definition:

$$z_{n+1} - x_{n+1} = \alpha_{n+1}(z_{n+1} - x_n^+) \Rightarrow \alpha_{n+1} = (\star) \tag{141}$$

As $B - A = \lambda(B - x_n^+)$ for some $\lambda \in \mathbb{R}$ (colinearity), with previous computation we have $\lambda = \alpha_{n+1}$ and then

$$B - A = \alpha_{n+1}(B - x_n^+) \tag{142}$$

Similarly, the colinearity of Bx_n^+ and $x_n^+ x_{n-1}^+$ together with $x_n x_n^+ \parallel v_n B$ leads to $B - x_n^+ = \lambda(x_n^+ - x_{n-1}^+)$ where

$$\lambda = \frac{\|B - x_n^+\|}{\|x_n^+ - x_{n-1}^+\|} = \frac{\|v_n - x_n\|}{\|x_n - x_{n-1}^+\|} \tag{143}$$

where we have

$$v_n - x_n = \beta_n z_n + (1-\beta_n) x_n - x_n = \beta_n(z_n - x_n) = \beta_n \alpha_n (z_n - x_{n-1}^+) \tag{144}$$

$$x_n - x_{n-1}^+ = (1-\alpha_n)(z_n - x_{n-1}^+) \tag{145}$$

such that

$$B - x_n^+ = \beta_n \frac{\alpha_n}{1-\alpha_n}(x_n^+ - x_{n-1}^+) \tag{146}$$

Combining (142) and (146), we get

$$A - x_n^+ = (B - x_n^+) - (B - A) = (B - x_n^+) = (B - x_n^+) - \alpha_{n+1}(B - x_n^+) \tag{147}$$

$$= (1-\alpha_{n+1})(B - x_n^+) = \beta_n \frac{(1-\alpha_{n+1})\alpha_n}{1-\alpha_n}(x_n^+ - x_{n-1}^+) \tag{148}$$

Then, we study $x_{n+1} - A$. As $Ax_{n+1} \parallel Bz_{n+1}$ we have by Thales theorem

$$\frac{\|x_{n+1} - A\|}{\|z_{n+1} - B\|} = \frac{\|x_{n+1} - x_n^+\|}{\|z_{n+1} - x_n^+\|} = 1 - \alpha_{n+1} \tag{149}$$

Last equality because $x_{n+1} - x_n^+ = (1 - \alpha_{n+1})(z_{n+1} - x_n^+)$. We then have

$$x_{n+1} - A = (1 - \alpha_{n+1})(z_{n+1} - B) = (1 - \alpha_{n+1})((z_{n+1} - v_n) - (B - v_n)) \tag{150}$$

Where we recall $v_n = \beta_n z_n + (1 - \beta_n)x_n$. We have:

$$z_{n+1} - v_n = (x_n^+ - x_n)\frac{\eta_n}{s} \tag{151}$$

Then we use $x_n x_n^+ \parallel v_n B$ to us Thales theorem once again:

$$\frac{\|B - v_n\|}{\|x_n^+ - x_n\|} = \frac{\|v_n - x_{n-1}^+\|}{\|x_n - x_{n-1}^+\|} := (\star\star) \tag{152}$$

We have $v_n - x_{n-1}^+ = \beta_n z_n + (1 - \beta_n)x_n - x_{n-1}^+$. We have also

$$\begin{aligned}
x_n &= \alpha_n x_{n-1}^+ + (1 - \alpha_n)z_n && (153)\\
\Rightarrow &\beta_n z_n - \beta_n x_n = \alpha_n \beta_n (z_n - x_{n-1}^+) && (154)\\
\Rightarrow &\beta_n z_n + (1 - \beta_n)x_n - x_{n-1}^+ = \alpha_n \beta_n (z_n - x_{n-1}^+) + x_n - x_{n-1}^+ && (155)\\
\Rightarrow &V_n - x_{n-1}^+ = (\frac{\alpha_n \beta_n}{1 - \alpha_n} + 1)(x_n - x_{n-1}^+) && (156)
\end{aligned}$$

This induces that $(\star\star) = (\frac{\alpha_n \beta_n}{1-\alpha_n} + 1)$ and then

$$B - v_n = \left(\frac{\alpha_n \beta_n}{1 - \alpha_n} + 1\right)(x_n^+ - x_n) \tag{157}$$

Finally, injecting (151) and (157) in (150), we get

$$\begin{aligned}
x_{n+1} - A &= (1 - \alpha_{n+1})\left(\frac{\eta_n}{s}(x_n^+ - x_n) - (\frac{\alpha_n \beta_n}{1 - \alpha_n} + 1)(x_n^+ - x_n)\right) && (158)\\
&= (1 - \alpha_{n+1})\left(\frac{\eta_n}{s} - \frac{\alpha_n \beta_n}{1 - \alpha_n} + 1\right)(x_n^+ - x_n) && (159)
\end{aligned}$$

We can conclude by combining (147) and (158) that

$$x_{n+1} = x_n^+ + \beta_n \frac{(1 - \alpha_{n+1})\alpha_n}{1 - \alpha_n}(x_n^+ - x_{n-1}^+) + (1 - \alpha_{n+1})\left(\frac{\eta_n}{s} - \frac{\alpha_n \beta_n}{1 - \alpha_n} + 1\right)(x_n^+ - x_n) \tag{160}$$

□

Corollary 2. *The algorithm 1 with parameters $s \leqslant \frac{1}{L}$, $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$, $\beta_n = 1 - \gamma\sqrt{\mu s}$ and $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$ can be written as the following 2 points scheme*

$$\begin{cases} y_n = x_n + \frac{1-\gamma\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_n - x_{n-1}) + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\gamma - 1)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla F(y_n) \end{cases} \tag{161}$$

Proof. We just apply previous result with $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$, $\beta_n = 1 - \gamma\sqrt{\mu s}$ et $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$. □

C.1.3 Proof of theorem 3

In this section, we detail the proof of Theorem 3 whose statement is recalled here: let $F = f + g$ where $f : \mathbb{R}^d \to \mathbb{R}$ is a L-smooth function for some $L > 0$ and $g : \mathbb{R}^d \to \mathbb{R}$ is convex, proper, lower semi-continuous. Assume that F has a non empty set of minimizers and that f is $(1, \mu)$-strongly quasar convex with respect to $x_F^* \in \arg\min\ F$ and (ρ, L)-curvatured for some $\rho \leqslant L$. Let $(x_n)_{n\in\mathbb{N}}$ be generated by Algorithm 2:

$$\begin{aligned} y_n &= \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} &= \operatorname{prox}_{sg}(y_n - s\nabla f(y_n)) := T_s(y_n) \\ z_{n+1} &= \beta_n z_n + (1-\beta_n) y_n - \frac{\eta_n}{s}(y_n - T_s(y_n)) \end{aligned}$$

with parameters

$$s \leqslant \frac{1}{L}, \quad \alpha_n = \frac{1}{1+\sqrt{\mu s}} := \alpha, \quad \beta_n = 1 - \gamma\sqrt{\mu s} := \beta, \quad \eta_n = \frac{\sqrt{s}}{\sqrt{\mu}} := \eta.$$

Assuming that $\rho \geqslant -\sqrt{\frac{\mu}{s}}$ we want to prove that:

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant 2\left(1 - \sqrt{\mu s}\right)^n \left(F(x_0) - F^*\right) \tag{162}$$

where $F^* = \min\ F$. Let x^* be the unique minimizer of F. As for the proof of Theorem 2, we introduce the following Lyapunov energy:

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|z_n - x^*\|^2 \tag{163}$$

and we seek the parameters et conditions for which the following inequality holds:

$$E_{n+1} - E_n \leqslant cE_n \tag{164}$$

with $c < 0$ being as small as possible. We will then deduce the convergence rate (162) by induction.

Proof. The proof is very similar to Theorem 2. Let us consider the same Lyapunov energy:

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|z_n - x_F^*\|^2 \tag{165}$$

Step 1. We start by calculating $E_{n+1} - E_n$, and the exact same computations as in Step 1 for Theorem 2 leads to:

$$\begin{aligned} E_{n+1} - E_n = &-(1-\beta)E_n + F(x_{n+1}) - F^* - \beta\left(F(x_n) - F^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x_F^*\|^2 \\ &+ \frac{\mu}{2s^2}\eta^2\|y_n - T_s(y_n)\|^2 - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \\ &+ \frac{\mu\eta}{s}\langle y_n - T_s(y_n), x_F^* - y_n\rangle + \frac{\alpha\beta\eta\mu}{(1-\alpha)s}\langle y_n - T_s(y_n), x_n - y_n\rangle. \end{aligned} \tag{166}$$

just replacing the gradient $\nabla f(y_n)$ by the composite gradient $\frac{1}{s}(y_n - T_s(y_n))$, where $T_s(y_n) := \operatorname{prox}_{sg}(y_n - s\nabla f(y_n))$.

Step 2. Let us now prove that for any $n \in \mathbb{N}$, we have: $E_{n+1} - E_n \leqslant -(1-\beta)E_n$ for some well-chosen values of the parameters β, η and α.

To control the scalar products, first note that:

$$2\langle y_n - T_s(y_n), x_F^* - y_n\rangle = \|T_s(y_n) - x_F^*\|^2 - \|y_n - T_s(y_n)\|^2 - \|y_n - x_F^*\|^2. \tag{167}$$

Combining the prox-grad inequality ([7, Theorem 10.16]): for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$F(x) - F(T_s(y)) \geqslant \frac{1}{2s}\|x - T_s(y)\|^2 - \frac{1}{2s}\|x - y\|^2 + f(x) - f(y) - \langle\nabla f(y), x - y\rangle, \qquad \text{(Prox-Grad)}$$

applied at $x = x_F^*$ and $y = y_n$, and the definition of strong quasar convexity in the sense of Definition 6:

$$\forall n \in \mathbb{N},\ f(x_F^*) - f(y_n) - \langle \nabla f(y_n), x_F^* - y_n \rangle \geqslant \frac{\mu}{2}\|x_F^* - y_n\|^2,$$

we prove that for any $n \in N$,

$$\begin{aligned} 2s\left(F^* - F(T_s(y_n))\right) &\geqslant \|x_F^* - T_s(y_n)\|^2 - \|x_F^* - y_n\|^2 + 2s\left(f(x_F^*) - f(y_n) - \langle \nabla f(y_n), x_F^* - y_n\rangle\right) \\ &\geqslant \|x_F^* - T_s(y_n)\|^2 - \|x_F^* - y_n\|^2 + \mu s\|x_F^* - y_n\|^2. \end{aligned}$$

Hence, reinjecting into (167) and remembering $x_{n+1} = T_s(y_n)$, we get:

$$\forall n \in \mathbb{N},\ \langle y_n - T_s(y_n), x_F^* - y_n\rangle \leqslant -s\left(F(x_{n+1}) - F^*\right) - \frac{\mu s}{2}\|y_n - x_F^*\|^2 - \frac{1}{2}\|y_n - x_{n+1}\|^2. \tag{168}$$

Similarly, consider then the second scalar product:

$$\begin{aligned} 2\langle x_n - y_n,\ y_n - T_s(y_n)\rangle &= \|T_s(y_n) - x_n\|^2 - \|y_n - T_s(y_n)\|^2 - \|y_n - x_n\|^2 && (169) \\ &\leqslant 2s\left(F(x_n) - F(T_s(y_n))\right) + 2s(f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle) \\ &\quad -\|y_n - T_s(y_n)\|^2 && (170) \\ &\leqslant 2s\left(F(x_n) - F(x_{n+1})\right) + 2s(f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle) \\ &\quad -\|y_n - x_{n+1}\|^2 && (171) \end{aligned}$$

using again (Prox-Grad) evaluated at $x = x_n$ and $y = y_n$ and $x_{n+1} = T_s(y_n)$.

Reinjecting (168) and (171) into the expression of $E_{n+1} - E_n$ obtained at the end of Step 1, we get:

$$\begin{aligned} E_{n+1} - E_n \leqslant\ & -(1-\beta)E_n + \left(1 - \mu\eta - \mu\beta\frac{\alpha}{1-\alpha}\eta\right)\left(F(x_{n+1}) - F^*\right) + \beta\left(\frac{\alpha\eta\mu}{1-\alpha} - 1\right)\left(F(x_n) - F^*\right) \\ & + \frac{\mu}{2}\left(1 - \beta - \mu\eta\right)\|y_n - x_F^*\|^2 + \frac{\mu\eta}{2s}\left(\frac{\eta}{s} - 1 - \frac{\alpha\beta}{1-\alpha}\right)\|y_n - x_{n+1}\|^2 \\ & - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\left(f(x_n) - f(y_n) - \langle \nabla f(y_n), y_n - x_n\rangle\right) \end{aligned}$$

As for Theorem 2, choose: $\eta = \frac{\sqrt{s}}{\sqrt{\mu}}$, $\beta = 1 - \eta\mu = 1 - \sqrt{\mu s}$ and $\alpha = \frac{1}{1+\eta\mu} = \frac{1}{1+\sqrt{\mu s}}$ to cancel out the terms in $F(x_{n+1}) - F^*$, $\|y_n - x_F^*\|^2$, $F(x_n) - F^*$ and $\|y_n - x_{n+1}\|^2$. We then get:

$$E_{n+1} - E_n \leqslant -(1-\beta)E_n + \beta\left(f(y_n) + \langle \nabla f(y_n), x_n - y_n\rangle - f(x_n)\right) - \beta\frac{\sqrt{\mu}}{2\sqrt{s}}\|x_n - y_n\|^2. \tag{172}$$

Finally, assuming additionally that f is (ρ, L)-curvatured for some $\rho \leqslant L$, observe that:

$$\forall n \in \mathbb{N}, f(y_n) + \langle \nabla f(y_n), x_n - y_n\rangle - f(x_n) \leqslant -\frac{\rho}{2}\|x_n - y_n\|^2,$$

which induces:

$$\forall n \in \mathbb{N},\ E_{n+1} - E_n \leqslant -(1-\beta)E_n - \frac{\beta}{2}\left(\rho + \frac{\sqrt{\mu}}{\sqrt{s}}\right)\|x_n - y_n\|^2$$

Provided that $\rho \geqslant -\sqrt{\frac{\mu}{s}}$, we finally obtain the expected inequality, namely: $E_{n+1} - E_n \leqslant -\sqrt{\mu s}E_n$ for all $n \in \mathbb{N}$, and we can conclude the proof exactly the same way as for Theorem 2, Step 3. □

D Continuous analysis through High resolution ODEs

Derivation of ODE Recall that the algorithm we prove convergence in Theorem 2 can be written

$$\begin{cases} y_n = x_n + \frac{1-\gamma\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_n - x_{n-1}) + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\gamma - 1)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla F(y_n) \end{cases} \tag{173}$$

Writting only with respect to y_n, we get

$$y_{n+1} = y_n + \frac{1-\gamma\sqrt{\mu s}}{1+\sqrt{\mu s}}(y_n - y_{n-1}) - s\left(1 + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\gamma - 1)\right)\nabla F(y_n) - s\frac{1-\gamma\sqrt{\mu s}}{1+\sqrt{\mu s}}(\nabla F(y_n) - \nabla F(y_{n-1})) \tag{174}$$

The following development will be very close to the one introduced in [36]. We assume there exists a smooth curve X such that $X(t_n) = y_n$, where $t_n = n\sqrt{s}$. By Taylor development, we have

$$y_{n+1} = X(t_{n+1}) = X(t_n) + \sqrt{s}\dot{X}(t_n) + \frac{s}{2}\ddot{X}(t_n) + \frac{\sqrt{s}^3}{6}\dddot{X}(t_n) + \mathcal{O}(s^2) \tag{175}$$

$$y_{n-1} = X(t_{n-1}) = X(t_n) - \sqrt{s}\dot{X}(t_n) + \frac{s}{2}\ddot{X}(t_n) - \frac{\sqrt{s}^3}{6}\dddot{X}(t_n) + \mathcal{O}(s^2) \tag{176}$$

Another Taylor development gives

$$\nabla F(y_n) - \nabla F(y_{n-1}) = \nabla^2 F(X(t_n))\dot{X}(t_n)\sqrt{s} + \mathcal{O}(s) \tag{177}$$

Multiplying both sides of (174) by $\frac{1+\sqrt{\mu s}}{1-\gamma\sqrt{\mu s}}\frac{1}{s}$, we get

$$\frac{y_{n+1} + y_{n-1} - 2y_n}{s} + \frac{(1+\gamma)\sqrt{\mu s}}{1-\gamma\sqrt{\mu s}}\frac{y_{n+1} - y_n}{s} + \nabla F(y_n) - \nabla F(y_{n-1}) + \frac{1+\gamma\sqrt{\mu s}}{1-\gamma\sqrt{\mu s}}\nabla F(y_{n-1}) = 0 \tag{178}$$

Using Taylor developments above, we have

$$\ddot{X}(t_n) + \mathcal{O}(s) + \frac{(1+\gamma)\sqrt{\mu}}{1-\gamma\sqrt{\mu s}}\left[\dot{X}(t_n) + \frac{1}{2}\ddot{X}(t_n)\sqrt{s} + \mathcal{O}(s)\right] \tag{179}$$

$$+ \nabla^2 F(X(t_n))\dot{X}(t_n)\sqrt{s} + \mathcal{O}(s) + \left(\frac{1+\gamma\sqrt{\mu s}}{1-\gamma\sqrt{\mu s}}\right)\nabla F(X(t_n)) = 0 \tag{180}$$

Multiplying both sides by $1 - \gamma\sqrt{\mu s}$ and ignoring $\mathcal{O}(s)$ terms, we get that (174) is a discretization of the following ODE

$$(1 + \frac{1-\gamma}{2}\sqrt{\mu s})\ddot{X}(t) + (1+\gamma)\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 F(X(t))\dot{X} + (1+\gamma\sqrt{\mu s})\nabla F(X(t)) = 0 \quad \text{(NAG-SQC-ODE)}$$

Proposition. *Let F be $(\gamma, \mu)-$strongly quasar convex and L-smooth. Assume X is solution of (NAG-SQC-ODE) with $0 < s \leqslant \frac{1}{L}$, $X(0) = X_0$ and $\dot{X}(0) = 0$. Then:*

$$F(X(t)) - F^* \leqslant K_0(\gamma, \mu, L, s)\frac{1}{\gamma}(F(X_0) - F^*)e^{-\gamma\frac{\sqrt{\mu}}{2}t} \tag{181}$$

where $K_0(\gamma, \mu, L, s) \leqslant 7$.

Proof. We rewrite the ODE (NAG-SQC-ODE) the following way.

$$\upsilon\ddot{X}(t) + (1+\gamma)\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 F(X(t))\dot{X}(t) + (1+\gamma\sqrt{\mu s})\nabla F(X(t)) = 0 \tag{182}$$

Where $\upsilon = 1 + \frac{1-\gamma}{2}\sqrt{\mu s}$. Set the following Lyapunov function:

$$\mathcal{E}(t) = \delta(F(X(t)) - F^*) + \frac{1}{2}\|\upsilon\dot{X}(t) + \lambda(X(t) - x^*) + \sqrt{s}\nabla F(X(t))\|^2 \tag{183}$$

To lighten the following computations, we write $X(t)$ as X, and we do the same for the first and second derivatives of X. We have

$$\dot{\mathcal{E}}(t) = \delta\langle\dot{X}, \nabla F(X)\rangle + \langle\upsilon\dot{X}(t) + \lambda(X(t) - x^*) + \sqrt{s}\nabla F(X(t)), \upsilon\ddot{X} + \lambda\dot{X} + \sqrt{s}\nabla^2 F(X)\dot{X}\rangle \tag{184}$$

Injecting (NAG-SQC-ODE), we get

$$\dot{\mathcal{E}}(t) = \delta\langle \dot{X}, \nabla F(X)\rangle + \langle \upsilon\dot{X}(t) + \lambda(X(t) - x^*) \tag{185}$$

$$+ \sqrt{s}\nabla F(X(t)), (\lambda - (1+\gamma)\sqrt{\mu})\sqrt{\mu}\dot{X} - (1+\gamma\sqrt{\mu s})\nabla F(X)\rangle \tag{186}$$

$$= \delta\langle \dot{X}, \nabla F(X)\rangle + \upsilon(\lambda - (1+\gamma)\sqrt{\mu})(\|\dot{X}\|^2 \tag{187}$$

$$- \upsilon(1+\gamma\sqrt{\mu s})\langle \dot{X}, \nabla F(X)\rangle + \lambda(\lambda - (1+\gamma)\sqrt{\mu})\langle X - x^*, \dot{X}\rangle \tag{188}$$

$$- \lambda(1+\gamma\sqrt{\mu s})\langle X - x^*, \nabla F(X)\rangle + (\lambda - (1+\gamma)\sqrt{\mu})\sqrt{s}\langle \nabla F, \dot{X}\rangle - \sqrt{s}(1+\gamma\sqrt{\mu s})\|\nabla F(X)\|^2 \tag{189}$$

We set $(\lambda - (1+\gamma)\sqrt{\mu}) = -\upsilon\gamma\sqrt{\mu}$. Then to cancel $\langle \dot{X}, \nabla F(X)\rangle$ terms, we set

$$\delta = (\upsilon(1+\gamma\sqrt{\mu s}) + \upsilon\gamma\sqrt{\mu s}) = \upsilon(1+2\gamma\sqrt{\mu s}) \tag{190}$$

Then we get

$$\dot{\mathcal{E}}(t) \leqslant -\gamma\sqrt{\mu}\left(\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda}{\gamma\sqrt{\mu}}(1+\gamma\sqrt{\mu s})\langle X - x^*, \nabla F(X)\rangle + s\|\nabla F(X)\|^2\right) \tag{191}$$

$$- \sqrt{s}\|\nabla F(X)\|^2 \tag{192}$$

We use strong quasar convexity.

$$\dot{\mathcal{E}}(t) \leqslant -\gamma\sqrt{\mu}\left(\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda}{\sqrt{\mu}}(1+\gamma\sqrt{\mu s})(F(X) - F^*)\right. \tag{193}$$

$$\left. + \frac{\lambda\sqrt{\mu}}{2}(1+\gamma\sqrt{\mu s})\|X - x^*\|^2 + s\|\nabla F(X)\|^2\right) - \sqrt{s}\|\nabla F(X)\|^2 \tag{194}$$

$$= -\gamma\sqrt{\mu}\left(\frac{1}{2}\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda^2}{2}\|X - x^*\|^2\right) \tag{195}$$

$$\left. + \frac{\upsilon}{2}(1+2\gamma\sqrt{\mu s})(F(X) - F^*) + s\|\nabla F(X)\|^2 - \sqrt{s}\|\nabla F(X)\|^2\right) \tag{196}$$

$$- \gamma\sqrt{\mu}\left(\frac{\lambda\sqrt{\mu}}{2}(1+\gamma\sqrt{\mu s}) - \frac{\lambda^2}{2}\right)\|X - x^*\|^2 \tag{197}$$

$$- \gamma\sqrt{\mu}\left(\frac{\lambda}{\sqrt{\mu}}(1+\gamma\sqrt{\mu s}) - \frac{\upsilon}{2}(1+2\gamma\sqrt{\mu s})\right)(F(X) - F^*) - \gamma\frac{\sqrt{\mu}\upsilon^2}{2}\|\dot{X}\|^2 \tag{198}$$

We have $\frac{1}{2}\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda^2}{2}\|X - x^*\|^2 = \|\upsilon\dot{X} + \lambda(X - x^*)\|^2$. Then, we use:

$$\frac{1}{2}\|\upsilon\dot{X} + \lambda(X - x^*) + \sqrt{s}\nabla F(X)\|^2 \leqslant \|\upsilon\dot{X} + \lambda(X - x^*)\|^2 + s\|\nabla F(X)\|^2 \tag{199}$$

$$\Rightarrow -\frac{1}{4}\|\upsilon\dot{X} + \lambda(X - x^*) + \sqrt{s}\nabla F(X)\|^2 \geqslant -\frac{1}{2}\|\upsilon\dot{X} + \lambda(X - x^*)\|^2 - \frac{s}{2}\|\nabla F(X)\|^2 \tag{200}$$

This leads to:

$$\dot{\mathcal{E}}(t) \leqslant -\gamma\frac{\sqrt{\mu}}{2}\mathcal{E}(t) - \sqrt{s}\|\nabla F(X)\|^2 - \gamma\sqrt{\mu}\left(\frac{\lambda\sqrt{\mu}}{2}(1+\gamma\sqrt{\mu s}) - \frac{\lambda^2}{2}\right)\|X - x^*\|^2 \tag{201}$$

$$- \gamma\sqrt{\mu}\left(\frac{\lambda}{\sqrt{\mu}}(1+\gamma\sqrt{\mu s}) - \frac{\upsilon}{2}(1+2\gamma\sqrt{\mu s})\right)(F(X) - F^*) - \gamma\frac{\sqrt{\mu}\upsilon^2}{2}\|\dot{X}\|^2 \tag{202}$$

We have to check that some terms are negatives. We have $\lambda = \sqrt{\mu}(1+\gamma(1-\upsilon))$, and

$$\frac{\lambda\sqrt{\mu}}{2}(1+\gamma\sqrt{\mu s}) - \frac{\lambda^2}{2} = \frac{\lambda}{2}\left(\sqrt{\mu}(1+\gamma\sqrt{\mu s}) - \lambda\right) = \frac{\lambda\mu}{2}\left(\gamma\sqrt{\mu s} - \gamma(1-\upsilon)\right) = \frac{\lambda\gamma\mu\sqrt{s}}{2}\left(\frac{1-\gamma}{2}\right) \geqslant 0 \tag{203}$$

and

$$\frac{\lambda}{\sqrt{\mu}}(1+\gamma\sqrt{\mu s})-\frac{v}{2}(1+2\gamma\sqrt{\mu s})=\left(1-\frac{\gamma(1-\gamma)}{2}\sqrt{\mu s}\right)(1+\gamma\sqrt{\mu s})-\frac{1}{2}(1+\frac{1-\gamma}{2}\sqrt{\mu s})(1+2\gamma\sqrt{\mu s}) \tag{204}$$

$$=1+\gamma\sqrt{\mu s}-\frac{1}{2}(1+2\gamma\sqrt{\mu s})-\frac{\gamma(1-\gamma)}{2}\sqrt{\mu s}(1+\gamma\sqrt{\mu s}) \tag{205}$$

$$-\frac{1}{2}\frac{1-\gamma}{2}\sqrt{\mu s}(1+2\gamma\sqrt{\mu s}) \tag{206}$$

$$=\frac{1}{2}-\frac{1-\gamma}{2}\sqrt{\mu s}\left(\gamma(1+\gamma\sqrt{\mu s})+\frac{1}{2}+\gamma\sqrt{\mu s}\right) \tag{207}$$

We want to be sure that this quantity is positive. To do so, we will maximize the right term with respect to $\gamma\in[0,1]$. First, note that supposing $s\leqslant\frac{1}{L}$, we have:

$$\frac{1-\gamma}{2}\sqrt{\mu s}\left(\gamma(1+\gamma\sqrt{\mu s})+\frac{1}{2}+\gamma\sqrt{\mu s}\right)\leqslant\frac{1-\gamma}{2}\left(\gamma(1+\gamma)+\frac{1}{2}+\gamma\right)=\frac{1}{4}(1-\gamma)(1+4\gamma+2\gamma^2):=g(\gamma) \tag{208}$$

We have

$$4g(\gamma)=(1+3\gamma-2\gamma^2-2\gamma^3) \tag{209}$$

We now want to find critical points of g.

$$4g'(\gamma)=3-4\gamma-6\gamma^2 \tag{210}$$

We calculate the discriminant $\Delta=4^2+4*6*3=88$, inducing that the roots of g' are

$$x_1=-\frac{4+2\sqrt{22}}{12},\quad x_2=\frac{-4+2\sqrt{22}}{12} \tag{211}$$

We clearly have $x_1<0$. x_2 however belongs to $[0,1]$. We evaluate numerically $g(x_2)\approx 0.44<\frac{1}{2}$. We conclude that for all $\gamma\in[0,1]$ we have

$$\frac{1}{2}-\frac{1-\gamma}{2}\sqrt{\mu s}\left(\gamma(1+\gamma\sqrt{\mu s})+\frac{1}{2}+\gamma\sqrt{\mu s}\right)>0 \tag{212}$$

All the out of parenthesis terms are negative, so we conclude that:

$$\dot{\mathcal{E}}(t)\leqslant-\gamma\frac{\sqrt{\mu}}{2}\mathcal{E}(t)\Rightarrow\mathcal{E}(t)\leqslant\mathcal{E}(0)e^{-\gamma\frac{\sqrt{\mu}}{2}t} \tag{213}$$

Supposing $t_0=0$

Deducing rate on $F(X(t))-F^*$ Using initial conditions, we have

$$\mathcal{E}(0)=\delta(F(X_0)-F^*)+\frac{1}{2}\|\lambda(X_0-x^*)-\sqrt{s}\nabla F(X_0)\|^2 \tag{214}$$

$$\leqslant\delta(F(X_0)-F^*)+\lambda^2\|X_0-x^*\|^2+s\|\nabla F(X_0)\|^2 \tag{215}$$

$$\leqslant\left(\delta+\frac{2\lambda^2(2-\gamma)}{\gamma\mu}+2Ls\right)(F(X_0)-F^*) \tag{216}$$

Where the third inequality uses that F is $\frac{\mu\gamma}{2-\gamma}$-quadratic growth (Propostion 1) and that F is L-Smooth. Then, we have

$$F(X(t))-F^*\leqslant\underbrace{\left(\gamma+\frac{2\lambda^2(2-\gamma)}{\mu\delta}+2\gamma Ls\right)}_{:=K_0(\gamma,\mu,L,s)}\frac{1}{\gamma}(F(X_0)-F^*)e^{-\gamma\frac{\sqrt{\mu}}{2}t} \tag{217}$$

We need to check that $K_0(\gamma,\mu,L,s)$ is uniformly bounded. Note first that has $0<s\leqslant\frac{1}{L}$, we have

$$sL\leqslant 1,\quad \mu s\leqslant 1 \tag{218}$$

We bound now $\upsilon, \delta, \lambda$, that we already fixed in the proof.

$$1 \leq \upsilon := 1 + \frac{1-\gamma}{2}\sqrt{\mu s} \leqslant \frac{3}{2} \tag{219}$$

$$1 \leqslant \delta := \upsilon(1 + 2\gamma\sqrt{\mu s}) \leqslant \frac{9}{2} \tag{220}$$

$$\sqrt{\mu}\left(1 - \frac{1}{8}\right) \leq \lambda := \sqrt{\mu}\left(1 + \frac{\gamma(\gamma-1)}{2}\sqrt{\mu s}\right) \leqslant \sqrt{\mu} \tag{221}$$

In the last inequality, we used the well known fact that $0 \leqslant p(1-p) \leqslant \frac{1}{4}$, for all $p \in [0,1]$.
We thus can explicitly compute that

$$K_0(\gamma, \mu, L, s) \leqslant (1 + 4 + 2) = 7 \tag{222}$$

□

D.1 Computation of derivation difference

We show how we can slightly modify the proof of Theorem 2 in order to exhibit the derivation difference mentionned in section 4.2. We start from equation (131) from the proof of Theorem 2.

$$\begin{aligned} E_{n+1} - E_n = &-(1-\beta)E_n + F(x_{n+1}) - F^* - \beta\left(F(x_n) - F^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 + \frac{\mu}{2}\eta^2\|\nabla F(y_n)\|^2 \\ &- \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle\nabla F(y_n), y_n - x_n\rangle - \mu\eta\langle\nabla F(y_n), y_n - x^*\rangle. \end{aligned} \tag{223}$$

As in the original proof, we use (129)-(130), the inequality given by the assumptions over the class of functions. The only difference here will be that we use the L-smooth property to bound $F(x_{n+1}) - F^*$ instea d of $\|\nabla F(y_n)\|^2$, *i.e.* we use:

$$\forall s \leqslant \frac{1}{L},\ \forall n \in \mathbb{N},\ F(x_{n+1}) \leqslant F(y_n) - \frac{s}{2}\|\nabla F(y_n)\|^2 \tag{224}$$

We then get:

$$E_{n+1} - E_n \leqslant -(1-\beta)E_n + (1 - \gamma\mu\eta)(F(y_n) - F^*) + (\frac{\mu}{2}\eta^2 - \frac{s}{2})\|\nabla F(y_n)\|^2 \tag{225}$$

$$- \beta(F(x_n) - F^*) - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle\nabla F(y_n), y_n - x_n\rangle - \frac{\mu\beta(1-\beta)}{2}\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \tag{226}$$

$$+ \frac{\mu}{2}(1 - \beta - \gamma\eta\mu)\|y_n - x^*\|^2 \tag{227}$$

Recall the choices of parameter of Theorem 2, that are:

$$s \leqslant \frac{1}{L}, \quad \alpha_n = \frac{1}{1 + \sqrt{\mu s}} := \alpha, \quad \beta_n = 1 - \gamma\sqrt{\mu s} := \beta, \quad \eta_n = \frac{\sqrt{s}}{\sqrt{\mu}} := \eta.$$

Using this choice of parameters, we get the following equation:

$$\begin{aligned} E_{n+1} - E_n \leqslant &-\gamma\sqrt{\mu s}E_n + (1 - \gamma\sqrt{\mu s})\left(F(y_n) - F(x_n) + \langle\nabla F(y_n), x_n - y_n\rangle\right) \\ &- \gamma(1 - \gamma\sqrt{\mu s})\sqrt{\frac{\mu}{s}}\|y_n - x_n\|^2. \end{aligned} \tag{228}$$

This is exactly equation (44).