Accelerated Mirror Descent for Non-Euclidean Star-convex Functions

Clement Lezane CLEMENT.LEZANE@UNIV-TOULOUSE.FR
ANITI, Toulouse, France

Sophie Langer S.LANGER@RUB.DE
Faculty for Mathematics, Ruhr University, Bochum, Germany

Wouter M. Koolen WMKOOLEN@CWI.NL
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Acceleration for non-convex functions is a fundamental challenge in optimization. We revisit star-convex functions, which are strictly unimodal on all lines through a minimizer. Hinder et al. (2020) accelerate unconstrained star-convex minimization of functions that are smooth with respect to the Euclidean norm. To do so, they add a certain binary search step to gradient descent. In this paper, we accelerate unconstrained star-convex minimization of functions that are *weakly* smooth with respect to an *arbitrary* norm. We add a binary search step to mirror descent, generalize the approach and refine its complexity analysis. We prove that our algorithms have sharp convergence rates for star-convex functions with α-Hölder continuous gradients and demonstrate that our rates are nearly optimal for p-norms.

1. Introduction

Accelerated gradient descent by Nesterov (1983) has proven to be a powerful tool for first-order optimization, significantly improving algorithmic performance. Nowadays several extensions of accelerated methods exist including coordinate descent methods (Nesterov, 2012; Fercoq and Richtárik, 2015), distributed gradient descent (Qu and Li, 2020) and proximal methods (Frostig et al., 2015) with applications across widespread domains such as image deblurring (Beck and Teboulle, 2009), compressed sensing (Becker et al., 2011) and deep learning (Sutskever et al., 2013).

Theoretical results analyzing accelerated methods mainly focus on smooth convex optimization, demonstrating optimal convergence and improved performance against traditional gradient methods, see Nesterov (1983), Bubeck (2015) and d'Aspremont et al. (2018) for further references. However, these assumptions are violated in modern machine learning applications due to their non-convex energy landscapes. Attempts to understand acceleration for non-convex functions are made, for instance, by leveraging the Łojasiewicz inequalities (Bolte et al., 2010). Further classes of non-convex functions are discussed by Carmon et al. (2017) and Hinder et al. (2020)[Appendix A.1]

In Nesterov and Polyak (2006) the convexity assumption is relaxed to so-called star-convex functions, a class of structured non-convex functions that are strictly unimodal on all lines through a minimizer. While convex functions exhibit unimodality along *all* lines, star-convexity is a strictly weaker condition. We emphasize that star-convexity differs from quasi-convexity as introduced in Boyd and Vandenberghe (2004); the two notions are not hierarchically related (Appendix L). Lee and Valiant (2016) provide examples of functions that are star-convex but not convex. Additionally, Zhou et al. (2019b) and Kleinberg et al. (2018) present evidence that neural network loss functions often

are star-convex (but not convex), emphasizing its relevance for modern machine learning applications. Notably, Zhang et al. (2023) and Zhang et al. (2025) show that convexifying loss functions can improve the performance of iterative models, with the resulting loss functions exhibiting star-convex properties.

For star-convex functions, Hinder et al. (2020) present an accelerated framework using binary search techniques. Their method is nearly optimal for objective functions having a Lipschitz gradient with respect to the Euclidean norm. However, recent machine learning results such as by Yang et al. (2015); Adil et al. (2019) minimize the objective functions in *non*-Euclidean norms, which are not covered by the findings of Hinder et al. (2020). This gap inspires us to extend and generalize their algorithm to arbitrary normed spaces. We combine a generalized "line-search" style algorithm (Hinder et al., 2020, Algorithm 2) with accelerated mirror descent to minimize star-convex functions with α-Hölder continuous gradients (also known as weakly smooth functions, see Definition 3). Contrary to common belief, our new framework reveals that acceleration depends on a norm's regularity (Fact 12), not the objective function's convexity.

We demonstrate that in case of p-norms, our algorithm is nearly optimal for $p > 1$. Our approach further extends to the L_1 norm though with an additional dimensionality-dependent cost tied to challenges in constructing an appropriate so-called distance generating function. By showing that non-convex functions can be minimized using accelerated algorithms in non-Euclidean geometry, we broaden the conventional understanding of which functions admit acceleration. Our findings suggest that a wider class of non-convex functions can be efficiently minimized, potentially driving further research in this direction.

Notation. Throughout this paper, we use $\|\cdot\|$ to denote an arbitrary norm and write $\|\cdot\|_*$ for the associated dual norm. We denote a minimizer of a function $F : \mathbb{R}^d \to \mathbb{R}$ by $x_\star$ and we say that a point is ϵ-optimal if $F(x) \leq F(x_\star) + \epsilon$. We denote the scalar product by $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$. We say that for two functions f and g, $f(x) = \mathcal{O}(g(x))$ if there exist positive real numbers M and x_0 such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$. Further, $f(x) = \Omega(g(x))$ if $\limsup_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| > 0$.

1.1. Literature review

Previous results on star-convex functions primarily focus on non-accelerated algorithms. Gower et al. (2020) and Hardt et al. (2016) show convergence guarantees for stochastic gradient descent, Nesterov and Polyak (2006) analyze cubic regularization and Lee and Valiant (2016) investigate a cutting plane method. Joulani et al. (2017) apply the star-convexity assumption to stochastic and online learning settings, showing regret bounds for first-order methods, leading to a $\mathcal{O}(1/T)$ convergence rate (with T being the number of gradient steps) in case that either the objective function or its gradient are Lipschitz.

Acceleration on star-convex functions has been studied by, e.g., Guminov and Gasnikov (2018) for low-dimensional subspace optimization and by Zhou et al. (2019a) for an one-dimensional line search-type of algorithm, both showing convergence rates to an ϵ-optimal point. However, as noted by Hinder et al. (2020), these algorithms can be limited by their potentially high computational costs for minimising general star-convex functions.

Hinder et al. (2020) propose an accelerated framework using binary search, showing near-optimal performance for τ-star-convex functions (see Definition 1) that are L-smooth (see Definition 3) with respect to the Euclidean norm in a domain bounded by radius R. Similar to Guminov and Gasnikov

(2018), their algorithm finds an ϵ-optimal solution in $\mathcal{O}(\epsilon^{-1/2}L^{1/2}\tau R\log(LR^2/\epsilon))$ iterations. For smooth star-convex functions, they also provide a lower bound of $\Omega(L^{1/2}\tau R\epsilon^{-1/2})$.

While the accelerated framework of Hinder et al. (2020) is limited to star-convex functions that are smooth with respect to the Euclidean norm, Nesterov's accelerated framework (Elster, 1993; Nemirovski and Nesterov, 1985) for convex functions applies to general normed spaces under a weaker smoothness assumption (Definition 3 below).

1.2. Our contributions

In this paper, we generalize the accelerated framework introduced by Hinder et al. (2020) to encompass weakly smooth star-convex functions with respect to a general norm. To address the challenges arising from the non-convexity of the objective function and the non-Euclidean geometry of the norm, we present a new convergence proof by using a regularity property of norms, involving novel arguments for bounded iterates to quantify convergence. Our key contributions are summarized as follows:

Acceleration for arbitrary norms. To extend the acceleration framework of Hinder et al. (2020) beyond Euclidean norms to general norms, we examine regularity properties of the derivative of the squared norm $\|\cdot\|^2$. We show that star-convexity and weak smoothness suffice to apply our generalized accelerated framework. Our results underscore that the key to acceleration lies in the convexity and homogeneity of norms, rather than convexity of the objective function.

Extended binary search. We generalize the binary search algorithm introduced by Hinder et al. (2020), originally designed for functions that are smooth with respect to the Euclidean norms, to encompass objective functions that are weakly smooth under arbitrary norms. We provide a streamlined convergence proof (Theorem 7) for our generalized algorithm and show that the number of oracle calls required for our binary search step is at most logarithmic. A key challenge in this setting is that our iterates may be unbounded, unlike in the Euclidean case. Additionally, we provide a lower bound for the binary search step when applied to smooth non-star-convex functions.

Nearly optimal performance for p-norm. For p-norms, we prove the same convergence rates as shown for convex functions by Nemirovski and Nesterov (1985), extending them to star-convex functions. According to the lower bounds by Guzmán and Nemirovski (2015), our rates are nearly optimal, up to a logarithmic factor. Table 1 contextualizes our results in the existing literature.

The article is organized as follows. In Section 2, we outline the problem setting and provide preliminaries. Section 3 presents our generalized accelerated algorithm and discusses its convergence guarantees for various settings. Section 4 introduces our binary search sub-algorithm and its running time. In Section 5 we put the two ingredients together and obtain our main bound. A discussion of future research directions is given in Section 6. All proofs are provided in the Appendix.

2. Problem setting and preliminaries

In this paper we consider the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^d} F(x)$$

with the unknown function $F \in \mathcal{F}$ coming from a given class $\mathcal{F} \subset [\mathbb{R}^d \to \mathbb{R}]$ of differentiable functions. We briefly discuss constrained domains in Section 6. We assume a first-order access

p	Bound	Convex	τ-star-convex
$(1,2]$	Upper	$\mathcal{O}_\kappa\left(\left(\frac{LR^\kappa}{\epsilon}\right)^{\frac{2}{3\kappa-2}}\right)$ Nemirovski and Nesterov (1985)	$\mathcal{O}_\kappa\left(\left(\frac{L\tau^2R^\kappa}{\epsilon}\right)^{\frac{2}{3\kappa-2}}\log^2\left(\frac{L\tau R}{\epsilon}\right)\right)$ **Our work** (Corollary 11(i))
	Lower	$\Omega_\kappa\left(\left(\frac{LR^\kappa}{\epsilon[\ln d]^{\kappa-1}}\right)^{\frac{2}{3\kappa-2}}\right)$ Guzmán and Nemirovski (2015)	
$[2,\infty]$	Upper	$\mathcal{O}_{p,\kappa}\left(\left(\frac{LR^\kappa}{\epsilon}\right)^{\frac{p}{\kappa p+\kappa-p}}\right)$ Nemirovski and Nesterov (1985)	$\mathcal{O}_{p,\kappa}\left(\left(\frac{L\tau^pR^\kappa}{\epsilon}\right)^{\frac{p}{\kappa p+\kappa-p}}\log^2\left(\frac{L\tau R}{\epsilon}\right)\right)$ **Our work** (Corollary 11(ii))
	Lower	$\Omega_{p,\kappa}\left(\left(\frac{LR^\kappa}{\epsilon}\right)^{\frac{p}{\kappa p+\kappa-p}}\right)$ Guzmán and Nemirovski (2015)	

Table 1: Summary of the relation between existing work and our results for τ-star-convex functions. R is the radius of the domain of the τ-star-convex objective function, (L,κ) is its smoothness with respect to a p-norm $\|\cdot\|_p$ with $p>1$ and ϵ the accuracy of the algorithm. Since convex functions are 1-star-convex, the lower bounds from Guzmán and Nemirovski (2015) apply to both.

model, meaning that the algorithm has access to the objective function as well as its gradient. The performance of the algorithm is measured by comparing (after T iterations) the function value of its output against $\min_{x\in\mathbb{R}^d} F(x)$. In the following, we assume that $\mathcal{F}$ describes the class of star-convex, (weakly) smooth functions with respect to a given (not necessarily Euclidean) norm $\|\cdot\|$. The formal definitions for this function class are introduced next.

2.1. Definitions

We consider the class of τ-star-convex functions, as introduced by Hardt et al. (2016), which are defined as follows.

Definition 1 (τ-star-convexity) *A continuously differentiable function $F:\mathbb{R}^d\to\mathbb{R}$ is said to be* τ-star-convex *for $\tau>0$ if there exists $x_\star\in\arg\min F$ such that for all $x\in\mathbb{R}^d$*

$$\tau\langle\nabla F(x), x-x_\star\rangle \;\geq\; F(x)-F(x_\star). \tag{1}$$

Note that, all convex functions are 1-star-convex (Joulani et al., 2017). The condition becomes weaker as τ grows; conversely, if allowed, 0-star convex function would be identically constant. In our unconstrained optimization setting, we observe that $x_\star$ must be a global minimizer (as the gradient vanishes at global minima).

For first-order methods, it is often assumed that the gradient is bounded, Lipschitz or α-Hölder continuous (see e.g. Nemirovski and Yudin, 1983; Diakonikolas and Guzmán, 2024), implying that the objective function satisfies certain smoothness conditions. In the following, we introduce the definition of (weak) smoothness with respect to a norm $\|\cdot\|$ using the notion of Bregman divergence, defined as follows.

Definition 2 (Bregman divergence) *For a continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, the* Bregman divergence *from $x \in \mathbb{R}^d$ to $y \in \mathbb{R}^d$ is defined by*

$$D_F(x, y) := F(x) - F(y) - \langle \nabla F(y), x - y \rangle.$$

Note that we do not assume convexity of F, and consequently we may have $D_F(x, y) < 0$.

Definition 3 (Weak smoothness) *A continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be* (L, κ)-weakly smooth *with respect to norm $\|\cdot\|$ for $1 < \kappa \leq 2$ if for all $x, y \in \mathbb{R}^d$*

$$|D_F(x, y)| \leq \frac{L}{\kappa} \|x - y\|^\kappa.$$

The special case $\kappa = 2$ is simply referred to as smooth.

To develop our mirror descent algorithm, we use a so-called distance-generating function (see Assumption 1), which fulfils a strong or uniform convexity assumption with respect to an arbitrary norm, defined as follows.

Definition 4 (Uniform convexity) *A continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be* (μ, q)-uniformly convex *with respect to norm $\|\cdot\|$ for $q \geq 2$ if for all $x, y \in \mathbb{R}^d$*

$$D_F(x, y) \geq \frac{\mu}{q} \|x - y\|^q.$$

The special case $q = 2$ is called strongly convex.

The classic acceleration setting considers $q = \kappa = 2$. In general, we have $q > 2 > \kappa$.

3. Main results

Our accelerated optimization algorithm for weakly smooth, star-convex functions incorporates both a mirror descent step and a proximal step, similar to d'Aspremont et al. (2018). However, as we will see in the proof sketch, our analysis exploits the specific structure of unconstrained optimization. We start by summarizing our assumptions.

Assumption 1 *We have a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ to be optimized, a norm $\|\cdot\|$ defined on $\mathbb{R}^d$, a function $\psi : \mathbb{R}^d \to \mathbb{R}$ customarily called the distance-generating function, an order $q \geq 2$ and constants τ, L, κ, μ such that*

- *F is τ-star-convex (Definition 1).*
- *F is (L, κ)-weakly smooth w.r.t. $\|\cdot\|$ (Definition 3).*
- *The distance-generating function ψ is (μ, q)-uniformly convex w.r.t. $\|\cdot\|$ (Definition 4).*

For an arbitrary norm, constructing a distance-generating function that satisfies the uniform convexity assumption of Definition 4 requires careful consideration. Below we present examples of distance-generating functions for p-norms and composite norms, each meeting the required conditions of uniform convexity.

Example 1 (p-norms) *To construct the corresponding distance-generating function, we define $\psi(x) = \frac{1}{q}\|x\|_p^q$, where the appropriate order q depends on p. Specifically, for $p \in (1, 2]$, we set $q = 2$, ensuring that ψ is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$ (see Blair, 1985; Juditsky and Nemirovski, 2008). For $p > 2$, we choose $q = p$, ensuring that ψ is $(2^{-\frac{p(p-2)}{p-1}}, p)$-uniformly convex with respect to $\|\cdot\|_p$ (see Zalinescu, 1983; d'Aspremont et al., 2024).*

For both cases, we will show sharp convergence guarantees of our algorithm, as summarized in Table 1.

Example 2 (Composite norms and distance generating function) *For composite norms, the distance-generating function is constructed by combining the distance-generating functions associated with the individual norms, ensuring the desired uniform convexity property. Consider, for instance, the following composite norm*

$$\|x\|_{2\circ 1.5} := \sqrt{\frac{1}{2}\|x_{1:d/2}\|_2^2 + \frac{1}{2}\|x_{(d/2+1):d}\|_{1.5}^2}. \tag{2}$$

By Example 1, $\psi_1(x_{1:d/2}) = \frac{1}{2}\sum_{i=1}^{d/2} x_i^2$ is 1-strongly convex with respect to $\|\cdot\|_2$ and $\psi_2(x_{d/2+1,d}) = \frac{1}{1.5}(\sum_{i=d/2+1}^{d}|x_i|^{1.5})^{2/1.5}$ is $\frac{1}{2}$-strongly convex with respect to $\|\cdot\|_{1.5}$. Hence, we define the composite distance-generating function as $\psi(x) := \frac{1}{2}\psi_1(x_{1:d/2}) + \frac{1}{2}\psi_2(x_{d/2+1,d})$. For any $x, y \in \mathbb{R}^d$, we denote the Bregman divergence by $D_\psi(x, y)$ and we note that

$$\begin{aligned} D_\psi(x,y) &= \frac{1}{2}\Big(\psi_1(x_{1:d/2}) - \psi_1(y_{1:d/2}) - \langle \nabla\psi_1(y_{1:d/2}), x_{1:d/2} - y_{1:d/2}\rangle\Big) \\ &\quad + \frac{1}{2}\Big(\psi_2(x_{(d/2+1):d}) - \psi_2(y_{(d/2+1):d}) - \langle \nabla\psi_2(y_{(d/2+1):d}), x_{(d/2+1):d} - y_{(d/2+1):d}\rangle\Big) \\ &\geq \frac{1}{4}\|x_{1:d/2} - y_{1:d/2}\|_2^2 + \frac{1}{8}\|x_{(d/2+1):d} - y_{(d/2+1):d}\|_{1.5}^2 \;\geq\; \frac{\|x-y\|_{2\circ 1.5}^2}{4}. \end{aligned}$$

We use the strong convexity of ψ_1 and ψ_2 in the last line. Thus, ψ is $\frac{1}{2}$-strongly convex with respect to the composite norm $\|\|_{2\circ 1.5}$. In general, the strong convexity coefficient of the composite norm is determined by the weaker of the two individual norms.

Our proposed method, outlined in Algorithm 1, extends the accelerated method of Hinder et al. (2020), which addresses the case of smoothness ($\kappa = 2$) for the Euclidean norm $\|\cdot\|_2$. This norm is known to be strongly convex with respect to itself, with $q = 2$ and $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. Similar to their method, ours maintains an auxiliary sequence $(x_t)_{t\geq 1}$ of iterates, from which it derives a sequence $(x_t^{md})_{t\geq 1}$ of gradient query points as well as the output sequence $(x_t^{ag})_{t\geq 1}$ of approximate minimizers. In each iteration, the gradient point x_t^{md} is found by binary search, which we extend from Hinder et al. (2020) to handle weak smoothness in Section 4. While this search is not necessary for convex F, it appears to be a reasonable trade-off for the relaxation to star-convexity, see also the discussion in Section 6. Hinder et al. (2020) update the iterates x_t and approximate minimizers x_t^{ag} each using gradient descent. However, to address a non-Euclidean norm, we instead employ a mirror descent step with respect to the Bregman divergence D_ψ to update x_t, and a proximal step with respect to $\frac{1}{q}\|\cdot\|^q$ to update x_t^{ag}. The accelerated order $q \geq 2$ depends on the norm $\|\cdot\|$ and will asymptotically determine the convergence speed of our algorithm, with a smaller q leading to faster rates.

Our algorithm takes as input the parameter schedules $(C_t, \epsilon_t, \alpha_t, \eta_t)$, with (α_t, η_t) being the step sizes and (C_t, ϵ_t) being the stopping conditions for the binary search. All tuned parameters are polynomial functions of the horizon T, and are given in Theorem 5 and Appendix C.

Algorithm 1 Accelerated Mirror Descent with Binary Search (for the setting of Assumption 1)

Require: Starting point $x_1 = x_1^{ag} \in \mathbb{R}^d$, iterations $T \geq 0$, parameters $(C_t, \epsilon_t, \alpha_t, \eta_t)_{t \geq 1}$

1: **For** $1 \leq t \leq T$ **do**

2: *Binary search* (Alg. 2) with endpoints (x_t, x_t^{ag}) and parameters (C_t, ϵ_t) to find $\lambda_t \in [0,1]$, $x_t^{md} = \lambda_t x_t^{ag} + (1-\lambda_t)x_t$ such that $\langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle + C_t\Big(F(x_t^{md}) - F(x_t^{ag})\Big) \leq \epsilon_t$

3: Compute $x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \{\eta_t \langle \nabla F(x_t^{md}), x \rangle + D_\psi(x, x_t)\}$

4: Compute $x_{t+1}^{ag} = \arg\min_{x \in \mathbb{R}^d} \{\alpha_t \langle \nabla F(x_t^{md}), x \rangle + \frac{\mu}{q}\|x - x_t^{md}\|^q\}$

5: **EndFor**

6: **Return** x_{T+1}^{ag}

As we will see in Theorem 7, the total number of values and gradients queried in each binary search step is $\mathcal{O}(\log T)$. In turn, the total number of gradients queried in T rounds is $\mathcal{O}(T \log T)$. We note that the optimal tuning of step sizes will depend on the number of iterations T.

3.1. Convergence result

In this section, we present the convergence result of our algorithm. The full proof is provided in Appendix C, with a proof sketch given in Section 3.2.

Theorem 5 *In the setting of Assumption 1 with $\kappa < q$, Algorithm 1 with the tuning below returns x_{T+1}^{ag} after T iterations with precision*

$$F(x_{T+1}^{ag}) - F_\star \;=\; \mathcal{O}_{q,\kappa}\left(\frac{L\tau^q}{T^{\frac{\kappa q + \kappa - q}{q}}}\left(\frac{D_\psi(x_\star, x_1)}{\alpha L} + \left(\frac{\alpha L}{\mu}\right)^{\frac{\kappa}{q-\kappa}} \log(T)\right)\right)$$

where $\mathcal{O}_{q,\kappa}$ omits constants depending only on (q, κ). This is achieved for any $\alpha > 0$ by

$$\alpha_t := \Big(\tau(q-\beta)\Big)^{q-\kappa}\frac{\alpha}{t^\beta}, \qquad \eta_t := \alpha_t\Big(\frac{t}{\tau(q-\beta)}\Big)^{q-1}.$$

$$C_t := \Big(\frac{\eta_t}{\alpha_t}\Big)^{q_*-1} - \frac{1}{\tau} = \frac{1}{\tau}\Big((q-\beta)t - 1\Big) \qquad \textit{where} \qquad \begin{cases} r & := \frac{q-\kappa}{\kappa} > 0 \\ M & := (\frac{r}{q})^r \geq 0 \\ \beta & = q - \kappa + \frac{q-\kappa}{q} \end{cases}$$

$$\epsilon_t := \frac{\alpha^{\frac{q}{q-\kappa}}}{t\eta_t}\frac{M^{1/r}}{\mu^{1/r}}L^{1+1/r}.$$

If q and κ are close to 2, our rate is close to the classic T^{-2} accelerated rate by Nemirovski and Nesterov (1985). The case $q = \kappa = 2$ requires special handling, as the above rate would yield an infinite exponent in the last term. We introduce a refined tuning for this case along with a tighter analysis in Appendix J.

Tuning the parameters. The general step sizes α_t and η_t are polynomial in the iteration number t with the main constraint being that their exponent must remain bounded. Specifically, we set $\alpha_t \sim \alpha t^{-\beta}$, where $\alpha > 0$ is a coefficient determined based on the Bregman divergence $D_\psi(x_\star, x_1)$. Additionally, η_t is chosen as $\eta_t \sim t^{1 - \frac{3q-\kappa(1+q)}{q}}$, which recovers the classic Lipschitz convex setting

$\eta_t \sim 1/\sqrt{t}$ for $q = 2, \kappa = 1$ and the Nesterov step size of $\eta_t \sim t$ for $q = \kappa = 2$. The choice of $C_t \sim t$ follows from the mirror descent analysis, while ϵ_t is chosen to decrease sufficiently fast such that $\eta_t \epsilon_t \lesssim 1/t$ sums to $\mathcal{O}(\log T)$. Full details are provided in Appendix C.

Tuning the parameter α. The optimal tuning for α in the preceding result depends on the radius $D_\psi(x_\star, x_1)$, which is unknown a priori. It also depends on $\log T$, while we prefer tuning independent of the time horizon T. To avoid both issues, we suggest the following practical tuning.

Corollary 6 *If a bound* $\frac{1}{\mu} D_\psi(x_\star, x_1) \leq B^q$ *is available, then the tuning of Theorem 5 with* $\alpha = \frac{\mu}{L}\left(\frac{(q-\kappa)B^q}{\kappa}\right)^{\frac{q-\kappa}{q}}$ *guarantees* $F(x^{ag}_{T+1}) - F_\star = \mathcal{O}_{q,\kappa}\left(L\tau^q B^\kappa \frac{\log(T)}{T^{\frac{\kappa q+\kappa-q}{q}}}\right)$.

If a bound on $\frac{1}{\mu} D_\psi(x_\star, x_1)$ is not available, it is possible to instead fix $B = 1$ and tune α accordingly. The convergence remains similar with $\max\{D_\psi(x_\star, x_1), 1\}$ replacing B. In the next sections, we will sketch the proof of Theorem 5 and examine the complexity of binary search in Section 4. After that, we will finish the complexity analysis of our algorithm in Section 5.

3.2. Proof Sketch of Theorem 5 (Convergence rate of Algorithm 1)

The proof can be decomposed into three steps. As the first step of the proof, we use star-convexity of the objective F to linearize it around the gradient point x^{md}_t. Together with the mirror descent iterates x_{t+1} from line 3 of Algorithm 1, we obtain the upper bound

$$\begin{aligned}\frac{\eta_t}{\tau}\Big(F(x^{md}_t) - F_\star\Big) \leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t\langle \nabla F(x^{md}_t), x^{md}_t - x_t\rangle \\ + \eta_t\langle \nabla F(x^{md}_t), x_t - x_{t+1}\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q.\end{aligned} \tag{3}$$

Next, we consider the unconstrained proximal step in line 4 of Algorithm 1, i.e.,

$$x^{ag}_{t+1} = \arg\min_{x\in\mathbb{R}^d} \Big\{\alpha_t\langle \nabla F(x^{md}_t), x\rangle + \frac{\mu}{q}\|x - x^{md}_t\|^q\Big\}.$$

This step leads to the following gap between consecutive accelerated potentials $\big(F(x^{md}_t), F(x^{ag}_{t+1})\big)$

$$\eta_t\langle \nabla F(x^{md}_t), x_t - x_{t+1}\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q \leq A_t\Big(F(x^{md}_t) - F(x^{ag}_{t+1})\Big) + B_t$$

with parameters A_t, B_t specified in (15). The term B_t is the classical residual after combining the weak smoothness assumption of F with the inexact gradient-trick (see Chen and Teboulle, 1993; d'Aspremont et al., 2018; d'Aspremont et al., 2024). This step was previously known to hold in the convex case (d'Aspremont et al., 2018) or in the unconstrained Euclidean case (Hinder et al., 2020) for different reasons. For the non-convex non-Euclidean unconstrained case, one needs to examine carefully the regularity of the norm square (see Appendix A).

After computing the new accelerated point x^{ag}_{t+1} from the middle point x_t, the previous objective bound (3) can be manipulated to bound the potential gap between two iterations $\big(F(x^{ag}_t), F(x^{ag}_{t+1})\big)$.

$$\begin{aligned}A_t\Big(F(x^{ag}_{t+1}) - F_\star\Big) \leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t\langle \nabla F(x^{md}_t), x^{md}_t - x_t\rangle \\ + \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x^{md}_t) - F(x^{ag}_t)\Big) + \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x^{ag}_t) - F_\star\Big) + B_t.\end{aligned} \tag{4}$$

As the last step, the choice of λ_t in line 2 of Algorithm 1 ensures that

$$\eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle + \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x_t^{md}) - F(x_t^{ag})\Big) \le \eta_t \epsilon_t.$$

By showing that our tuning satisfies $A_t - A_{t-1} \le \frac{\eta_t}{\tau}$, $\eta_t \epsilon_t = \frac{\mu}{t}$ and $B_t \approx \frac{1}{t}$, we get

$$A_t\Big(F(x_{t+1}^{ag}) - F_\star\Big) \le D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + A_{t-1}\Big(F(x_t^{ag}) - F_\star\Big) + O(1/t), \quad (5)$$

which telescopes to our result. To obtain the desired accelerated convergence rate in Table 1, we still need to account for the complexity of finding λ_t. In Section 4, we show that this additional cost remains only a logarithmic factor, even in the non-Euclidean case.

4. Binary search for Hölder smooth functions

In each round t, Algorithm 1 performs binary search to find the momentum parameter λ_t that satisfies

$$\lambda_t g'(\lambda_t) + C_t g(\lambda_t) \le \epsilon_t, \quad (6)$$

where we use the abbreviation $g(\lambda) \triangleq F(\lambda x_t^{ag} + (1-\lambda)x_t) - F(x_t^{ag})$ so that in particular $g(1) = 0$. The condition (6) is crucial for the convergence proof of Theorem 5. The core idea behind our Algorithm 2 is inspired by Algorithm 2 of Hinder et al. (2020). We aim to efficiently find a λ_t that satisfies (6) in the more general setting where F is weakly smooth with respect to a possibly non-Euclidean norm.

Note that if $g(0) \le 0$ or g is decreasing at $\lambda = 1$, then (6) is immediately satisfied. If neither of these conditions hold, then $g(0) > g(1)$ and $g'(1) > 0$, indicating that g must switch at least once from decreasing to increasing. For the largest λ_* fulfilling $g'(\lambda_*) = 0$ it holds that $g(\lambda_*) < 0$. We further show in the proof of Theorem 7 that all $\lambda \in [\lambda_* - \delta, \lambda_*]$, with a properly chosen δ, must satisfy condition (6). Next we quantify the efficiency of our algorithm in finding any λ in that interval.

Algorithm 2 Generalized binary search

Require: Objective function F, endpoints (x_t, x_t^{ag}), parameters $(C_t, \epsilon_t) \ge 0$
  Define $g(\lambda) \triangleq F(\lambda x_t^{ag} + (1-\lambda)x_t) - F(x_t^{ag})$
  if $g'(1) \le \epsilon_t$ **then return** $(\lambda_t = 1, x_t^{md} = x_t^{ag})$
  if $C_t g(0) \le \epsilon_t$ **then return** $(\lambda_t = 0, x_t^{md} = x_t)$
  Initialize $[a, b] = [0, 1]$
  loop
    Set $\lambda_t = \frac{a+b}{2}$
    if $g'(\lambda_t) + C_t g(\lambda_t) \le \epsilon_t$ **then return** $(\lambda_t, x_t^{md} = \lambda_t x_t^{ag} + (1-\lambda_t)x_t)$
    if $g(\lambda_t) > 0$ **then** $a = \lambda_t$ **else** $b = \lambda_t$ ▷ iterate with $[a, \lambda_t]$ or $[\lambda_t, b]$
  end loop

Theorem 7 *Under Assumption 1 (in particular F is (L, κ)-weakly smooth), and for fixed stopping parameters (C_t, ϵ_t) and endpoints (x_t, x_t^{ag}), Algorithm 2 finds λ, satisfying*

$$x_t^{md} = \lambda x_t^{ag} + (1-\lambda)x_t$$
$$\langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle + C_t\Big(F(x_t^{md}) - F(x_t^{ag})\Big) \le \epsilon_t$$

in $\mathcal{O}_\kappa\left(\max\{\log(C_t), \log\left(\frac{L\|x_t - x_t^{ag}\|^\kappa}{\epsilon_t}\right)\} + 1\right)$ *iterations.*

Hinder et al. (2020) analyze their binary search algorithm for smooth functions, i.e., specifically with $\kappa = 2$ in Definition 3. Our algorithm accommodates a general κ, addressing the broader class of weakly smooth functions. While the running time depends on κ, it appears only as an exponent in the logarithm. Consequently, this introduces only a constant multiplicative factor, thereby generalizing the algorithm of Hinder et al. (2020).

Proof Sketch. The proof can be roughly divided into *three* parts. We first show that there exists a λ_* satisfying the strengthening of condition (6) with $\epsilon_t = 0$. Next, we show that this implies that condition (6) holds for all $\lambda \in [\lambda_* - \delta, \lambda_*]$, where δ is properly chosen depending on the weak smoothness parameters (L, κ) of F and the parameters $(C_t, \epsilon_t, x_t, x_t^{ag})$ of the algorithm. Finally we prove that our algorithm finds a λ within that interval after at most $\mathcal{O}\Big(\log\left(\frac{LC_t\|x_t - x_t^{ag}\|^\kappa}{\epsilon_t}\right)\Big)$ iterations (details in Appendix E).

4.1. Polynomial growth of our iterates

To complete our analysis, we bound the iterates (x_t, x_t^{ag}) generated by Algorithm 1. In fact, our Theorem 7 shows that the number of iterations is bounded by a term logarithmic in t provided the distance $\|x_t - x_t^{ag}\|$ grows at most polynomially in t. This insight is formalized next.

Theorem 8 *Let $D_\psi(x_\star, x_1) \le B$. Under Assumption 1, suppose we run Algorithm 1 such that*

$$\text{for all } t \ge 1 \qquad \max\left(\left(\frac{qL\eta_t}{\kappa}\right)^{\frac{1}{q-1}}, \left(\frac{L\alpha_t}{\kappa}\right)^{\frac{1}{q-1}}\right) \le K^{n_1} t^{n_2},$$

where $n_1, n_2 \ge 0$, and $K^{n_1} > B$ with K^{n_1} potentially depending on (L, τ, μ). Define the iterate radius by $R_t = \max\{\|x_t - x_\star\|, \|x_t^{ag} - x_\star\|, \|x_t^{md} - x_\star\|\}$. Then, for all $t \ge 1$, we have

$$R_t = \mathcal{O}\Big(K^{\frac{n_1(q-1)}{q-\kappa}} t^{\frac{(q-1)(n_2+1)}{q-\kappa}}\Big).$$

4.2. Lower bound for Binary search

Although the following theorem does not contribute to our upper bounds, we analyze the binary search algorithm without assuming star-convexity to highlight the efficiency of Algorithm 2 under weak smoothness. This section shows that the number of iterations needed in Theorem 7 to find a suitable λ satisfying condition (6) is both sufficient and necessary - fewer itations could yield in a λ that does not satisfy condition (6). To show this lower bound, we determine the minimum number of iteration steps required to find a suitable λ for a univariate smooth target function.

Theorem 9 *Let $C, \epsilon, L_\star > 0$. Consider a sequence of N points $(\lambda_1, \dots, \lambda_N)$ evaluated at each iteration $(g(\lambda_i), g'(\lambda_i))$. If the number of iterations is insufficient meaning*

$$N < \log(5)\Big(\log\frac{L_\star}{\epsilon} + \log\frac{C}{(C+1)^2} - \log(88)\Big),$$

no algorithm can find λ fulfilling condition (6).

The theorem shows that if $N \lesssim \log(L_\star/\epsilon)$, it becomes impossible to determine λ with precision ϵ. The full version of this lower bound is given in Theorem 25.

5. Summary

By combining the convergence results from Corollary 6 with the complexity analysis presented in Theorem 7, we establish a bound on the precision error of Algorithm 1 in terms of the number of first-order oracle calls.

Corollary 10 *Let $T > 0$ be the number of iterations and $\frac{1}{\mu} D_\psi(x_\star, x_1) \leq B$. The tuning of Theorem 5 with $\alpha = \frac{\mu}{L}\left(\frac{(q-\kappa)B}{\kappa}\right)^{\frac{q-\kappa}{q}}$ guarantees $F(x_{T+1}^{ag}) - F_\star = \mathcal{O}_{q,\kappa}\left(L\tau^q B^{\frac{\kappa}{q}} \frac{\log(T)}{T^{\frac{\kappa q+\kappa-q}{q}}}\right)$. The algorithm's oracle usage over T iterations is upper bounded by $\mathcal{O}(T\log(LB\tau T))$, which represents the maximum number of times the oracle is called during the entire execution process.*

To apply Corollary 10 to the p-norms in Example 1, we consider a problem where both the initial points x_1 and minimizer $x_\star$ lie within a p-norm ball of radius R, i.e., $\|x_1\|_p, \|x_\star\|_p \leq R$, which implies that $\|x_1 - x_\star\|_p \leq 2R$ by the triangle inequality. For p-norms, recall that we use the distance-generating function $\psi(\cdot) = \frac{1}{q}\|\cdot\|_p^q$ with $q = \max\{2, p\}$. In Appendix H, we show that the Bregman divergence also satisfies $D_\psi(x_\star, x_1) \leq 2R^q$. Combining Theorems 5 and 8, we obtain the following convergence guarantees.

Corollary 11 *Consider the setting of Assumption 1 and let $\|\cdot\|_p$ denote the p-norm.*

(i) For $1 < p \leq 2$ and $\kappa < 2$, Algorithm 1 with $q = 2$ returns an ϵ-optimal solution within

$$\mathcal{O}_\kappa\left(\left(\frac{L\tau^2 R^\kappa}{\epsilon}\right)^{\frac{2}{3\kappa-2}} \log^2\left(\frac{L\tau R}{\epsilon}\right)\right)$$

evaluations of the function value and gradient.

(ii) For $p > 2$, Algorithm 1 with $q = p$ returns an ϵ-optimal solution within

$$\mathcal{O}_{p,\kappa}\left(\left(\frac{L\tau^p R^\kappa}{\epsilon}\right)^{\frac{p}{\kappa p+\kappa-p}} \log^2\left(\frac{L\tau R}{\epsilon}\right)\right)$$

evaluations of the function values and the gradients.

Note that the convergence guarantees of Corollary 11 for star-convex functions match the lower bounds proved for convex functions, completing Table 1. For $p = 1$, our analysis applies with a dimension factor, detailed in Appendix I. For the smooth case $\kappa = 2$ with $1 < p < 2$, where iterates may grow exponentially, we extend our algorithm while maintaining similar complexity; further details are provided in Appendix J.

6. Conclusion and future research

In this work, we introduce a novel class of structured non-convex functions, namely τ-star-convex functions that are weakly smooth with respect to an arbitrarily norm. Our framework applies to a broad subclass of star-convex functions and generalizes the setting analyzed in Hinder et al. (2020) which focuses on smooth star-convex functions with respect to the Euclidean norm.

We develop the accelerated Algorithm 1 to efficiently minimize any weakly smooth function within this broad class, building on and extending the binary search technique from Hinder et al.

(2020). Our analysis shows that the algorithm achieves a near-optimal accelerated convergence rate for all p-norms with $p > 1$. Additionally, our algorithm has a runtime dependence on $\tau^{\max\{2,p\}}$ for $p \geq 1$, shown to be optimal in the Euclidean case when $p = 2$, see Hinder et al. (2020).

While our work focuses on unconstrained minimization problems, future research could extend our approach to constrained settings. Currently, the proximal step for calculating x_{t+1}^{ag} in Algorithm 1 is shown to be effective only for unconstrained problems. For the constrained non-convex case, since evaluating the direction of the constrained proximal step becomes a challenge, it is unknown if accelerated convergence rates are possible. A more detailed elaboration on this is given in Appendix K.

Finally, our method operates within a first-order access model, meaning that the algorithm has access to both the objective function value and gradient at the evaluation points. It remains an open problem whether this can be extended to "pure" first-order access models, that rely solely on gradient evaluations. While this extension is feasible in the convex case, it is unclear whether it can be achieved for star-convex functions. Another promising direction is to explore the applicability of our method in a stochastic model. Although stochastic accelerated mirror descent has been previously analyzed in the literature (Hu et al., 2009), the main challenge lies in adapting the binary search algorithm to the noisy setting.

References

Deeksha Adil, Richard Peng, and Sushant Sachdeva. *Fast, provably convergent IRLS algorithm for p-norm linear regression*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542. URL https://doi.org/10.1137/080716542.

Stephen Becker, Jérôme Bobin, and Emmanuel J. Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011. doi: 10.1137/090756855. URL https://doi.org/10.1137/090756855.

Charles Blair. Problem complexity and method efficiency in optimization (A. S. Nemirovski and D. B. Yudin). *SIAM Review*, 27(2):264–265, 1985. doi: 10.1137/1027074. URL https://doi.org/10.1137/1027074.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010. ISSN 00029947. URL http://www.jstor.org/stable/25677828.

S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Number Teil 1 in Berichte über verteilte messysteme. Cambridge University Press, 2004. ISBN 9780521833783. URL https://books.google.de/books?id=mYm0bLd3fcoC.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. 2015. URL http://arxiv.org/abs/1405.4980.

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*, 2017. URL https://api.semanticscholar.org/CorpusID:35265682.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993. doi: 10.1137/0803026. URL https://doi.org/10.1137/0803026.

Alexandre d'Aspremont, Cristóbal Guzmán, and Martin Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018. doi: 10.1137/17M1116842. URL https://doi.org/10.1137/17M1116842.

Jelena Diakonikolas and Cristóbal Guzmán. Complementary composite minimization, small gradients in general norms, and applications. *Mathematical Programming*, Jan 2024. ISSN 1436-4646. doi: 10.1007/s10107-023-02040-5. URL https://doi.org/10.1007/s10107-023-02040-5.

Alexandre d'Aspremont, Cristóbal Guzmán, and Clément Lezane. Optimal algorithms for stochastic complementary composite minimization. *SIAM Journal on Optimization*, 34(1):163–189, 2024. doi: 10.1137/22M1530884. URL https://doi.org/10.1137/22M1530884.

K. H. Elster. Modern mathematical methods of optimization. 1993. URL https://api.semanticscholar.org/CorpusID:117469401.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015. doi: 10.1137/130949993. URL https://doi.org/10.1137/130949993.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, page 2540–2548. JMLR.org, 2015.

Robert Mansel Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, 2020. URL https://api.semanticscholar.org/CorpusID:219792513.

Sergey Guminov and Alexander V. Gasnikov. Accelerated methods for α-weakly-quasi-convex optimization problems. 2018. URL https://api.semanticscholar.org/CorpusID:121395144.

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2014.08.003. URL https://www.sciencedirect.com/science/article/pii/S0885064X14000831.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *J. Mach. Learn. Res.*, 19:29:1–29:44, 2016. URL https://api.semanticscholar.org/CorpusID:7597719.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/hinder20a.html.

Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/ec5aa0b7846082a2415f0902f0da88f2-Paper.pdf.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76, pages 681–720. PMLR, October 2017. URL https://proceedings.mlr.press/v76/joulani17a.html.

Anatoli Juditsky and Arkadi Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. arXiv, 2008. URL https://arxiv.org/abs/0809.0813.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kleinberg18a.html.

Jasper C.H. Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614, 2016. doi: 10.1109/FOCS.2016.71.

Arkadi Nemirovski and Yurii Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(85)90100-4. URL https://www.sciencedirect.com/science/article/pii/0041555385901004.

Arkadi Nemirovski and D.B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1983. ISBN 0471103454.

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. URL https://api.semanticscholar.org/CorpusID:145918791.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001. URL https://doi.org/10.1137/100802001.

Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006. URL https://api.semanticscholar.org/CorpusID:7964929.

Guannan Qu and Na Li. Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2020. doi: 10.1109/TAC.2019.2937496.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Jiyan Yang, Yinlam Chow, Christopher Ré, and Michael W. Mahoney. Weighted SGD for l_p regression with randomized preconditioning. *Proceedings of the ... Annual ACM-SIAM Symposium on Discrete Algorithms. ACM-SIAM Symposium on Discrete Algorithms*, 2016:558–569, 2015. URL https://api.semanticscholar.org/CorpusID:6360243.

W. H. Young. On classes of summable functions and their Fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87 (594):225–229, 1912. ISSN 09501207. URL http://www.jstor.org/stable/93236.

C. Zalinescu. On uniformly convex functions. *J. Math. Anal. Appl.*, 95:344–374, 1983. ISSN 0022-247X. doi: 10.1016/0022-247X(83)90112-9.

Yiqing Zhang, Xinming Huang, and Ziming Zhang. PRISE: demystifying deep lucas-kanade with strongly star-convex constraints for multimodel image alignment. *CoRR*, abs/2303.11526, 2023. doi: 10.48550/ARXIV.2303.11526. URL https://doi.org/10.48550/arXiv.2303.11526.

Ziming Zhang, Yuping Shao, Yiqing Zhang, Fangzhou Lin, Haichong Zhang, and Elke Rundensteiner. Deep loss convexification for learning iterative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1501–1513, 2025. doi: 10.1109/TPAMI.2024.3509860.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Accelerated primal-dual gradient descent with line search for convex, nonconvex, and nonsmooth optimization problems. *Doklady Mathematics*, 99, 2019a. URL https://doi.org/10.1134/S1064562419020042.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. *ArXiv*, abs/1901.00451, 2019b. URL https://api.semanticscholar.org/CorpusID:57373762.

Appendix A. A regularity property of norms

In this section, we will discuss about the regularity property for norms which is core for our acceleration scheme. For a norm $\|\cdot\|$ with associated dual norm $\|\cdot\|_*$, we define the gradient (or any sub-gradient) of the squared norm $\|\cdot\|^2$ by

$$\phi(x) \;:=\; \nabla\left(\frac{\|x\|^2}{2}\right) \;=\; \operatorname*{argmax}_{y:\|y\|_*=\|x\|} y^\intercal x.$$

Wherever the norm is not differentiable, we may pick $\phi(x)$ to be any sub-gradient or maximizer. To see the equality, recall that $\|x\| = \max_{z:\|z\|_*=1} z^\intercal x$ and let z attain that maximum. Then differentiation[1] gives $\phi(x) = \|x\| z$. Reparametrising by $y = \|x\| z$ gives the right hand side.

Fact 12 (Regularity of a norm) *For all $x \in \mathbb{R}^d$*

$$\begin{aligned}\langle \phi(x), x\rangle &= \|x\|^2,\\ \|\phi(x)\|_* &= \|x\|.\end{aligned}$$

These two properties are crucial in our accelerated framework (Algorithm 1), especially for weakly smooth objective functions. In particular, they allow for the choice of aggressive step-sizes and the derivation of accelerated convergence rates.

Lemma 13 *For $s \geq 1$, we define $\phi_s(x) := \nabla\left(\frac{\|x\|^s}{s}\right)$. For all $\alpha \geq 0$*

$$\begin{cases}\langle \phi_s(x), x\rangle &= \|x\|^s\\ \|\phi_s(x)\|_*^\alpha &= \|x\|^{\alpha(s-1)}.\end{cases}$$

Proof Recall that $\phi(x) = \nabla(\frac{\|x\|^2}{2})$. Then

$$\phi_s(x) = \nabla\Big(\frac{(\|x\|^2)^{s/2}}{s}\Big) = \frac{1}{s}\cdot\frac{s}{2}(\|x\|^2)^{\frac{s}{2}-1}\cdot\nabla\Big(\|x\|^2\Big) \;= (\|x\|^2)^{\frac{s}{2}-1}\cdot\frac{1}{2}\nabla\Big(\|x\|^2\Big) = \frac{\phi(x)}{\|x\|^{2-s}}.$$

As

$$\begin{aligned}\langle \phi_s(x), x\rangle &= \frac{\langle \phi(x), x\rangle}{\|x\|^{2-s}} = \frac{\|x\|^2}{\|x\|^{2-s}} = \|x\|^s,\\ \|\phi_s(x)\|_*^\alpha &= \frac{\|\phi(x)\|_*^\alpha}{\|x\|^{\alpha(2-s)}} = \frac{\|x\|^\alpha}{\|x\|^{\alpha(2-s)}} = \|x\|^{\alpha(s-1)},\end{aligned}$$

the assertion follows. ■

1. Recall that a sub-gradient of a maximum $x \mapsto \max_y f(x,y)$ is the derivative of the objective $\nabla_x f(x, y_*)$ with a maximizer y^* held fixed.

Appendix B. Useful lemma for convergence proof

Lemma 14 (Young's inequality, (Young, 1912)) *Let $a, b \geq 0$ and $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then it holds, that*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Lemma 15 *Let f, ν be convex functions and let ν be continuously differentiable. For*

$$u^\star = \arg\min_{u \in \mathcal{X}} \{f(u) + D^\nu(u, y)\},$$

and all u, it holds

$$f(u^\star) + D^\nu(u^\star, y) + D^f(u, u^\star) \leq f(u) + D^\nu(u, y) - D^\nu(u, u^\star).$$

Proof As $u^\star$ is a minimizer, it fulfills for all $u \in \mathcal{X}$ the first-order optimality condition

$$\langle \nabla f(u^\star) + \nabla D^\nu(u^\star, y), u - u^\star \rangle \geq 0, \tag{7}$$

with the gradient of the divergence D^ν taken with respect to the first argument. Applying the three-points identity Chen and Teboulle (1993), yields

$$\langle \nabla D^\nu(u^\star, y), u - u^\star \rangle = D^\nu(u, y) - D^\nu(u, u^\star) - D^\nu(u^\star, y),$$

leading to

$$f(u) - f(u^\star) - D^f(u, u^\star) = \langle \nabla f(u^\star), u - u^\star \rangle \geq D^\nu(u, u^\star) - D^\nu(u, y) + D^\nu(u^\star, y).$$

■

Lemma 16 *Set*

$$\begin{cases} r & := \frac{q-\kappa}{\kappa} > 0 \\ M & := (\frac{r}{q})^r \geq 0. \end{cases} \tag{8}$$

Then for all $\delta > 0, x, y \in \mathbb{R}^d$

$$\frac{\|x - y\|^\kappa}{\kappa} \leq \frac{M}{q\delta^r} \|x - y\|^{q} + \delta. \tag{9}$$

Proof Using Young's inequality 14 with $\frac{1}{a} + \frac{1}{b} = 1$, leads to $t \leq \frac{1}{az} t^a + \frac{1}{b} z^{b-1}$. By choosing $t := \frac{\|x-y\|^\kappa}{\kappa}$, $a = \frac{q}{\kappa}$, $b = \frac{q}{q-\kappa}$, $z = (b\delta)^{\frac{1}{b-1}} = (b\delta)^r$, it follows that

$$\frac{\|x - y\|^\kappa}{\kappa} \leq \frac{M}{q\delta^r} \|x - y\|^{q} + \delta,$$

where we use that

$$\frac{\kappa}{q} \left(\frac{q-\kappa}{q\delta}\right)^r \left(\frac{1}{\kappa}\right)^{q/\kappa} = \frac{1}{q\delta^r} \left(\frac{q-\kappa}{q\kappa}\right)^r = \frac{M}{q\delta^r}.$$

■

Appendix C. Convergence proof of Theorem 5

Our proof can be decomposed in *three* major steps. We first derive a bound for $F(x_t^{md}) - F_\star$ by analyzing our mirror descent step, then we proceed with an acceleration step comparing $F(x_{t+1}^{ag})$ and $F(x_t^{md})$ and finally we tune the parameters to satisfy the required conditions.

C.1. Mirror descent analysis

We apply Lemma 15 to the linear function $f(x) = \eta_t \langle \nabla F(x_t^{md}), x \rangle$ and set $\nu = \psi$, $u^\star = x_{t+1}$, $y = x_t$, $u = x$. For all $x \in \mathbb{R}^d$, this yields

$$\eta_t \langle \nabla F(x_t^{md}), x_{t+1} \rangle + D_\psi(x_{t+1}, x_t) \leq \eta_t \langle \nabla F(x_t^{md}), x \rangle + D_\psi(x, x_t) - D_\psi(x, x_{t+1}).$$

By setting $x = x_\star$ and subtracting $\eta_t \langle \nabla F(x_t^{md}), x_\star \rangle$ on both sides, we obtain

$$\eta_t \langle \nabla F(x_t^{md}), x_{t+1} - x_\star \rangle + D_\psi(x_{t+1}, x_t) \leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}).$$

We split

$$\eta_t \langle \nabla F(x_t^{md}), x_{t+1} - x_\star \rangle = \eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_\star \rangle + \eta_t \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle,$$

and use star-convexity of F

$$\frac{\eta_t}{\tau}\Big(F(x_t^{md}) - F_\star\Big) \leq \eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_\star \rangle$$

to find

$$\frac{\eta_t}{\tau}\Big(F(x_t^{md}) - F_\star\Big) + \eta_t \langle \nabla F(x_t^{md}), x_{t+1} - x_t^{md} \rangle + D_\psi(x_{t+1}, x_t) \leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}).$$

This leads to

$$\begin{aligned} &\frac{\eta_t}{\tau}\Big(F(x_t^{md}) - F_\star\Big) + D_\psi(x_{t+1}, x_t) \\ &\leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_{t+1} \rangle \\ &= D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle + \eta_t \langle \nabla F(x_t^{md}), x_t - x_{t+1} \rangle. \end{aligned} \tag{10}$$

By using that the distance-generating function ψ is (μ, q)-uniformly convex with respect to the norm $\|\cdot\|$, i.e., that it fulfills

$$D_\psi(x_{t+1}, x_t) \geq \frac{\mu}{q} \|x_{t+1} - x_t\|^q,$$

our equation becomes

$$\begin{aligned} &\frac{\eta_t}{\tau}\Big(F(x_t^{md}) - F_\star\Big) \\ &\leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle \\ &\quad + \eta_t \langle \nabla F(x_t^{md}), x_t - x_{t+1} \rangle - \frac{\mu}{q} \|x_{t+1} - x_t\|^q. \end{aligned} \tag{11}$$

The goal of the next step is to find an upper bound for

$$\eta_t \langle \nabla F(x_t^{md}), x_t - x_{t+1} \rangle - \frac{\mu}{q} \|x_{t+1} - x_t\|^q.$$

C.2. Choice of aggregated point

Recall from Algorithm 1 that

$$x_{t+1}^{ag} = \arg\min_{x\in\mathbb{R}^d}\{\alpha_t\langle\nabla F(x_t^{md}), x\rangle + \frac{\mu}{q}\|x - x_t^{md}\|^q\}$$

and define $\phi_q(x) := \nabla\Big(\frac{\|x\|^q}{q}\Big)$. For all $u \in \mathbb{R}^d$, it holds

$$\langle\alpha_t\nabla F(x_t^{md}) + \mu\phi_q(x_{t+1}^{ag} - x_t^{md}), u - x_{t+1}^{ag}\rangle \geq 0.$$

Choosing $u = x_{t+1}^{ag} + \lambda(x_{t+1} - x_t)$ for any $\lambda > 0$, leads to

$$\lambda\alpha_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle \leq \lambda\mu\langle\phi_q(x_{t+1}^{ag} - x_t^{md}), x_{t+1} - x_t\rangle.$$

Subtracting $\frac{\mu}{q}\|x_{t+1} - x_t\|^q$ on both sides, yields

$$\begin{aligned}
&\eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q \\
&\leq \frac{\mu\eta_t}{\alpha_t}\langle\phi_q(x_{t+1}^{ag} - x_t^{md}), x_{t+1} - x_t\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q \\
&= \mu\Big\langle\frac{\eta_t}{\alpha_t}\phi_q(x_{t+1}^{ag} - x_t^{md}), x_{t+1} - x_t\Big\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q
\end{aligned} \tag{12}$$

Let q_* be the convex conjugate of q, i.e., $\frac{1}{q} + \frac{1}{q_*} = 1$. A reformulation of Young's inequality in Lemma 14 gives

$$\langle a, b\rangle \leq |\langle a, b\rangle| \leq \|a\|_*\cdot\|b\| \leq \frac{\|a\|_*^{q_*}}{q_*} + \frac{\|b\|^q}{q},$$

leading to

$$\langle a, b\rangle - \frac{\|b\|^q}{q} \leq \frac{\|a\|_*^{q_*}}{q_*}.$$

Therefore (12) can be upper bounded by

$$\begin{aligned}
\mu\Big\langle\frac{\eta_t}{\alpha_t}\phi_q(x_{t+1}^{ag} - x_t^{md}), x_{t+1} - x_t\Big\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q &\leq \frac{\mu}{q_*}\Big\|\frac{\eta_t}{\alpha_t}\phi_q(x_{t+1}^{ag} - x_t^{md})\Big\|_*^{q_*} \\
&= \frac{\mu\eta_t^{q_*}}{q_*\alpha_t^{q_*}}\|\phi_q(x_{t+1}^{ag} - x_t^{md})\|_*^{q_*} \\
&= \frac{\mu\eta_t^{q_*}}{q_*\alpha_t^{q_*}}\|x_{t+1}^{ag} - x_t^{md}\|^{(q-1)q_*} \\
&= \frac{\mu\eta_t^{q_*}}{q_*\alpha_t^{q_*}}\|x_{t+1}^{ag} - x_t^{md}\|^q,
\end{aligned} \tag{13}$$

where the penultimate equality follows from Lemma 13. By setting $u = x_t^{md}$, we obtain from first-order optimality, that

$$\alpha_t\langle\nabla F(x_t^{md}), x_t^{md} - x_{t+1}^{ag}\rangle \geq \mu\langle\phi_q(x_{t+1}^{ag} - x_t^{md}), x_{t+1}^{ag} - x_t^{md}\rangle,$$

leading to

$$\begin{aligned}\alpha_t\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) &\geq \mu\|x_t^{md} - x_{t+1}^{ag}\|^q - \alpha_t D_F(x_{t+1}^{ag}, x_t^{md}) \\ &\geq \mu\|x_t^{md} - x_{t+1}^{ag}\|^q - \frac{L\alpha_t}{\kappa}\|x_t^{md} - x_{t+1}^{ag}\|^\kappa,\end{aligned}$$

where we use that F is weakly smooth for the last inequality. By choosing

$$\begin{aligned}r &:= \frac{q-\kappa}{\kappa} > 0 \\ M &:= \Big(\frac{r}{q}\Big)^r \geq 0,\end{aligned}$$

in Lemma 16, it follows for all $\delta_t > 0$

$$L\alpha_t \frac{\|x_t^{md} - x_{t+1}^{ag}\|^\kappa}{\kappa} \leq \frac{LM\alpha_t}{q\delta_t^r}\|x_t^{md} - x_{t+1}^{ag}\|^q + L\alpha_t\delta_t.$$

This leads to

$$\alpha_t\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) \geq \Big(\mu - \frac{LM\alpha_t}{q\delta_t^r}\Big)\|x_t^{md} - x_{t+1}^{ag}\|^q - L\alpha_t\delta_t.$$

For convenience, we set $\frac{LM\alpha_t}{q\delta_t^r} = \frac{\mu}{q}$ by choosing $\delta_t = \Big(\frac{L}{\mu}M\alpha_t\Big)^{1/r}$ and then get

$$\alpha_t\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) \geq \frac{\mu}{q_*}\|x_t^{md} - x_{t+1}^{ag}\|^q - \frac{M^{1/r}}{\mu^{1/r}}\Big(L\alpha_t\Big)^{(r+1)/r}.$$

Combining the previous equation with (12) and (13), yields

$$\begin{aligned}&\eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle - \frac{\mu}{q}\|x_{t+1} - x_t\|^q \\ \leq&\frac{\mu\eta_t^{q_*}}{\alpha_t^{q_*}}\frac{\|x_{t+1}^{ag} - x_t^{md}\|^q}{q_*} \\ \leq&\frac{\eta_t^{q_*}}{\alpha_t^{q_*}}\left(\alpha_t\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) + \frac{M^{1/r}}{\mu^{1/r}}\Big(L\alpha_t\Big)^{(r+1)/r}\right) \\ =&\frac{\eta_t^{q_*}}{\alpha_t^{q_*-1}}\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) + \frac{\eta_t^{q_*}}{\alpha_t^{q_*}}\frac{M^{1/r}}{\mu^{1/r}}\Big(L\alpha_t\Big)^{(r+1)/r}\end{aligned} \tag{14}$$

C.3. Simplified equation

To simplify notation, we will abbreviate

$$\begin{aligned}A_t &:= \frac{\eta_t^{q_*}}{\alpha_t^{q_*-1}} \\ B_t &:= \frac{\eta_t^{q_*}}{\alpha_t^{q_*}}\frac{M^{1/r}}{\mu^{1/r}}\Big(L\alpha_t\Big)^{1+1/r} = \frac{\eta_t^{q_*}}{\alpha_t^{q_*-1-1/r}}\frac{M^{1/r}}{\mu^{1/r}}L^{1+1/r}.\end{aligned} \tag{15}$$

Rearranging (11) and combining it with equation (14) leads to

$$\begin{aligned} &A_t\Big(F(x_{t+1}^{ag}) - F_\star\Big) \\ \leq &D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t\langle\nabla F(x_t^{md}), x_t^{md} - x_t\rangle \\ &+ \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x_t^{md}) - F(x_t^{ag})\Big) \\ &+ \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x_t^{ag}) - F_\star\Big) + B_t. \end{aligned} \tag{16}$$

With the binary search for λ_t, Algorithm 1 achieves

$$\eta_t\langle\nabla F(x_t^{md}), x_t^{md} - x_t\rangle + \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x_t^{md}) - F(x_t^{ag})\Big) \leq \eta_t\epsilon_t$$

within $\mathcal{O}(\log(\frac{L}{\epsilon_t}))$ iterations of function values and gradients. Thus, (16) becomes

$$\begin{aligned} A_t\Big(F(x_{t+1}^{ag}) - F_\star\Big) \leq &D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t\epsilon_t \\ &+ \Big(A_t - \frac{\eta_t}{\tau}\Big)\Big(F(x_t^{ag}) - F_\star\Big) + B_t. \end{aligned} \tag{17}$$

To derive the optimal convergence rate, we need the following conditions to be fulfilled for a fixed $\alpha > 0$:

$$\begin{cases} A_t - \frac{\eta_t}{\tau} & \leq A_{t-1} \\ B_t & \leq \alpha^{\frac{q}{q-\kappa}}\frac{1}{t}\frac{M^{1/r}}{\mu^{1/r}}L^{1+1/r}. \end{cases}$$

If the previous conditions hold, we can compute the telescopic sum in (17) and get

$$A_T\Big(F(x_{T+1}^{ag}) - F_\star\Big) \leq D_\psi(x_\star, x_1) + \mathcal{O}\Big(\alpha^{\frac{q}{q-\kappa}}\frac{L^{1+1/r}}{\mu^{1/r}}\log(T)\Big). \tag{18}$$

To find the optimal step-size, we choose

$$\beta = q - \kappa + \frac{q-\kappa}{q}, \quad \gamma \geq 0, \quad \Big(\frac{\eta_t}{\alpha_t}\Big)^{q_\star - 1} = \frac{1}{\tau(q-\beta)}t^{\gamma(q_\star - 1)}$$

and recall that $q - \beta = \frac{q\kappa + \kappa - q}{q} > 0$, $1 + \frac{1}{r} = 1 + \frac{\kappa}{q-\kappa} = \frac{q}{q-\kappa}$ and $\frac{q_*}{q} = \frac{1}{q-1}$. Then it follows

$$B_t \leq \alpha^{\frac{q}{q-\kappa}}\frac{1}{t}\frac{M^{1/r}}{\mu^{1/r}}L^{1+1/r},$$

leading to

$$\frac{\eta_t^{q_*}}{\alpha_t^{q_* - 1 - 1/r}}\frac{M^{1/r}}{\mu^{1/r}}L^{1+1/r} \leq \alpha^{\frac{q}{q-\kappa}}\frac{1}{t}\frac{M^{1/r}}{\mu^{1/r}}L^{1+1/r}$$

and thus

$$\alpha_t^{1+1/r} \leq \Big(\tau(q-\beta)\Big)^{\frac{q_\star}{q_\star - 1}}\frac{\alpha^{\frac{q}{q-\kappa}}}{t^{\gamma q_* + 1}} \Leftrightarrow \quad \alpha_t \leq \Big(\tau(q-\beta)\Big)^{q-\kappa}\frac{\alpha}{t^{\frac{\gamma(q-\kappa)}{q-1} + \frac{q-\kappa}{q}}}.$$

We further note that

$$A_t - A_{t-1} = \Big(\frac{\eta_t^{q_*}}{\alpha_t^{q_*-1}} - \frac{\eta_{t-1}^{q_*}}{\alpha_{t-1}^{q_*-1}}\Big) = \bigg(\eta_t\Big(\frac{\eta_t}{\alpha_t}\Big)^{q_\star-1} - \eta_{t-1}\Big(\frac{\eta_{t-1}}{\alpha_{t-1}}\Big)^{q_\star-1}\bigg) \leq \frac{\eta_t}{\tau}$$

is equialent to

$$\frac{A_t - A_{t-1}}{\eta_t} = \frac{1}{\tau(q-\beta)}\Big(t^{\gamma(q_*-1)} - \frac{\eta_{t-1}}{\eta_t}(t-1)^{\gamma(q_*-1)}\Big) \leq \frac{1}{\tau}.$$

Intuitively, choosing η_t as a polynomial will ensure that $\frac{\eta_{t-1}}{\eta_t}$ tends to 1. Fixing the parameters in the following way satisfies our condition

$$\begin{aligned}
\gamma &:= q-1 \\
\alpha_t &:= \Big(\tau(q-\beta)\Big)^{q-\kappa}\frac{\alpha}{t^\beta} \\
\eta_t &:= \alpha_t\Big(\frac{1}{\tau(q-\beta)}t^{\gamma(q_\star-1)}\Big)^{\frac{1}{q_\star-1}} = \alpha_t\Big(\frac{t}{\tau(q-\beta)}\Big)^{q-1}.
\end{aligned}$$

Note that $\gamma(q_*-1) = 1$. We verify that

$$\frac{\eta_{t-1}}{\eta_t} = \frac{\alpha_{t-1}}{\alpha_t}\Big(1-\frac{1}{t}\Big)^{q-1} = \Big(1-\frac{1}{t}\Big)^{q-1-\beta} = \Big(\frac{t-1}{t}\Big)^{q-1-\beta}$$

and

$$\begin{aligned}
\frac{A_t - A_{t-1}}{\eta_t} &= \frac{1}{\tau(q-\beta)}\Big(t - \frac{\eta_{t-1}}{\eta_t}(t-1)\Big) \\
&= \frac{1}{\tau(q-\beta)}\frac{1}{t^{q-1-\beta}}\Big(t^{q-\beta} - (t-1)^{q-\beta}\Big) \\
&= \frac{1}{\tau(q-\beta)}\frac{1}{t^{q-1-\beta}}\Big(\int_{t-1}^{t}(q-\beta)s^{q-\beta-1}ds\Big) \\
&\leq \frac{1}{\tau}\frac{1}{t^{q-\beta-1}}\cdot t^{q-1-\beta} = \frac{1}{\tau}
\end{aligned}$$

By choosing $\epsilon_t := \frac{B_t}{\eta_t}$ such that $\eta_t\epsilon_t \leq B_t$, this leads to

$$\begin{aligned}
A_t &= \frac{\eta_t^{q_*}}{\alpha_t^{q_*-1}} = \Big(\frac{\eta_t}{\alpha_t}\Big)^{q_*}\alpha_t = \Big(\frac{t}{\tau(q-\beta)}\Big)^{\frac{q_\star}{q_\star-1}}\alpha_t \\
&\propto \Big(\frac{t}{\tau}\Big)^q\cdot\frac{\alpha}{t^\beta} = \alpha t^{q-\beta}\cdot\frac{1}{\tau^q} \\
A_t &\propto \frac{\alpha}{\tau^q}t^{\frac{\kappa q+\kappa-q}{q}}
\end{aligned}$$

and finally shows

$$F(x_{T+1}^{ag}) - F_\star = \mathcal{O}_{q,\kappa}\Big(\frac{\tau^q D_\psi(x_\star, x_1)}{\alpha T^{\frac{\kappa q+\kappa-q}{q}}} + \frac{\alpha^{\frac{\kappa}{q-\kappa}}L^{\frac{q}{q-\kappa}}}{\mu^{\frac{\kappa}{q-\kappa}}}\frac{\tau^q\log(T)}{T^{\frac{\kappa q+\kappa-q}{q}}}\Big),$$

which proves Theorem 5.

Appendix D. Proof for bounded iterate

Theorem 17 *Let's assume that $D_\psi(x_\star, x_1) \leq B$. Under assumption 1, we also suppose that we are running Algorithm 1 with parameters tuned such that there are base K and exponents $n_1, n_2 \geq 0$ (possibly depending on (L, τ, μ)) ensuring $K^{n_1} > B$ and for all $t \geq 1$*

$$\max\left(\left(\frac{qL\eta_t}{\kappa}\right)^{\frac{1}{q-1}}, \left(\frac{L\alpha_t}{\kappa}\right)^{\frac{1}{q-1}}\right) \leq K^{n_1} t^{n_2}.$$

If we write $R_t = \max\{\|x_t - x_\star\|, \|x_t^{ag} - x_\star\|, \|x_t^{md} - x_\star\|\}$, then for all $t \geq 1$

$$R_t = \mathcal{O}\Big(K^{\frac{n_1(q-1)}{q-\kappa}} t^{\frac{(q-1)(n_2+1)}{q-\kappa}}\Big)$$

Proof We recall from our Algorithm 1 that

$$x_{t+1} = \arg\min_{x\in\mathbb{R}^d}\{\eta_t\langle\nabla F(x_t^{md}), x\rangle + D_\psi(x, x_t)\}$$
$$x_{t+1}^{ag} = \arg\min_{x\in\mathbb{R}^d}\{\alpha_t\langle\nabla F(x_t^{md}), x\rangle + \frac{\|x - x_t^{md}\|^q}{q}\}$$

Since our setting is unconstrained, the (sub-)gradients have to cancel :

$$\eta_t\nabla F(x_t^{md}) + \nabla\psi(x_{t+1}) - \nabla\psi(x_t) = 0$$
$$\alpha_t\nabla F(x_t^{md}) + \phi_q(x_{t+1}^{ag} - x_t^{md}) = 0$$

with $\psi(x) \propto \frac{\|x\|^q}{q}$ and $\phi_q(x) = \nabla\Big(\frac{\|x\|^q}{q}\Big)$. By the weak smoothness property of F:

$$\|\nabla F(x_t^{md})\|_* \leq \frac{L}{\kappa}\|x_t^{md} - x_\star\|^{\kappa-1}.$$

Additionally, from the uniform convexity of ψ and regularity of $\|\cdot\|$, it follows that:

$$\frac{\|x_{t+1} - x_t\|^q}{q} \leq D_\psi(x_t, x_{t+1}) + D_\psi(x_{t+1}, x_t) = \langle\nabla\psi(x_{t+1}) - \nabla\psi(x_t), x_{t+1} - x_t\rangle.$$

Applying the optimality condition for x_{t+1}, this becomes

$$\frac{\|x_{t+1} - x_t\|^q}{q} \leq \eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle$$

and by applying Cauchy-Schwartz, we obtain:

$$\frac{\|x_{t+1} - x_t\|^q}{q} \leq \eta_t\|\nabla F(x_t^{md})\|_* \times \|x_t - x_{t+1}\|,$$

where $\|\cdot\|_*$ is the dual norm associated with $\|\cdot\|$. Similarly for $x_{t+1}^{ag} - x_t^{md}$, we use the definition of ϕ_q and obtain from Lemma 13 that

$$\|x_{t+1}^{ag} - x_t^{md}\|^q = \langle\phi_q(x_{t+1}^{ag} - x_t^{md}), x_{t+1}^{ag} - x_t^{md}\rangle$$
$$\leq \alpha_t\|\nabla F(x_t^{md})\|_* \times \|x_{t+1}^{ag} - x_t^{md}\|.$$

By plugging in these results in the previous conditions, we derive the following bounds:

$$\|x_{t+1} - x_t\|^{q-1} \leq \frac{qL\eta_t}{\kappa}\|x_t^{md} - x_\star\|^{\kappa-1}$$
$$\|x_{t+1}^{ag} - x_t^{md}\|^{q-1} \leq \frac{L\alpha_t}{\kappa}\|x_t^{md} - x_\star\|^{\kappa-1}.$$

From the triangle inequality, we then obtain :

$$\begin{aligned}\|x_{t+1} - x_\star\| &\leq \|x_t - x_\star\| + \|x_{t+1} - x_t\| \\ &\leq \|x_t - x_\star\| + \Big(\frac{qL\eta_t}{\kappa}\Big)^{\frac{1}{q-1}}\|x_t^{md} - x_\star\|^{\frac{\kappa-1}{q-1}}\end{aligned}$$

and

$$\|x_{t+1}^{ag} - x_\star\| \leq \|x_t^{md} - x_\star\| + \Big(\frac{L\alpha_t}{\kappa}\Big)^{\frac{1}{q-1}}\|x_t^{md} - x_\star\|^{\frac{\kappa-1}{q-1}}.$$

Since x_{t+1}^{md} is a middle point, we can also deduce that

$$\begin{aligned}\|x_{t+1}^{md} - x_\star\| &= \|\lambda_{t+1}(x_{t+1}^{ag} - x_\star) + (1-\lambda_{t+1})(x_{t+1} - x_\star)\| \\ &\leq \lambda_{t+1}\|x_{t+1}^{ag} - x_\star\| + (1-\lambda_{t+1})\|x_{t+1} - x_\star\| \\ &\leq \max\Big(\|x_{t+1}^{ag} - x_\star\|, \|x_{t+1} - x_\star\|\Big)\end{aligned}$$

From the above inequality and using the assumption on (α_t, η_t) it follows that:

$$R_{t+1} \leq R_t + K^{n_1}t^{n_2}R_t^{\frac{\kappa-1}{q-1}}.$$

This recursive inequality shows that the growth is at most polynomial, implying that the sequence R_t cannot grow faster than a polynomial rate.

We will show by induction that $R_t \leq K^{S_t}t^{M_2}$, where $M_2 := \max\Big\{\frac{(q-1)(n_2+1)}{q-\kappa}, 1\Big\}$ and $S_t := n_1\sum_{i=0}^{t-1}\Big(\frac{\kappa-1}{q-1}\Big)^i$ is a geometric sum. First, we notice that

$$\begin{aligned}n_2 + 1 &\leq \frac{q-\kappa}{q-1}M_2 \\ \implies n_2 + M_2\frac{\kappa-1}{q-1} &\leq M_2 - 1 \\ \implies t^{n_2 + M_2\frac{\kappa-1}{q-1}} &\leq t^{M_2-1} \leq (t+1)^{M_2} - t^{M_2}.\end{aligned}$$

Then, by induction it then follows that

$$\begin{aligned}R_{t+1} \leq R_t + K^{n_1}t^{n_2}R_t^{\frac{\kappa-1}{q-1}} &\leq K^{S_t}t^{M_2} + K^{n_1 + S_t\times\frac{\kappa-1}{q-1}}t^{n_2 + M_2\frac{\kappa-1}{q-1}} \\ &\leq K^{S_{t+1}}\Big(t^{M_2} + t^{n_1 + M_2\frac{\kappa-1}{q-1}}\Big) \\ &\leq K^{S_{t+1}}(t+1)^{M_2}\end{aligned}$$

Since $S_t < \frac{n_1}{1-\frac{\kappa-1}{q-1}} = \frac{n_1(q-1)}{q-\kappa}$, we can simplify $R_t = \mathcal{O}\Big(K^{\frac{n_1(q-1)}{q-\kappa}}t^{\frac{(q-1)(n_2+1)}{q-\kappa}}\Big)$. ■

Appendix E. Proof for binary search

Before proving Theorem 7, we make two quick observations.

Lemma 18 *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a (L, κ)-weakly smooth function. Define $g : \mathbb{R} \to \mathbb{R}$ for $a, b \in \mathbb{R}^d$ by*

$$g(\lambda) = F(a + \lambda b).$$

Then for any λ_1, λ_2, we have

$$|g'(\lambda_1) - g'(\lambda_2)| \leq \frac{2L}{\kappa}\|b\|^\kappa |\lambda_1 - \lambda_2|^\kappa.$$

Proof We recall that :

$$g'(\lambda) = \langle \nabla F(a + \lambda b), b\rangle$$

and for any λ_1, λ_2, by weak smoothness:

$$-\frac{2L}{\kappa}\|\lambda_1 b - \lambda_2 b\|^\kappa \leq D_F(a + \lambda_1 b, a + \lambda_2 b) + D_F(a + \lambda_2 b, a + \lambda_1 b)) \leq \frac{2L}{\kappa}\|\lambda_1 b - \lambda_2 b\|^\kappa$$
$$-\frac{2L}{\kappa}\|\lambda_1 b - \lambda_2 b\|^\kappa \leq \langle \nabla F(a + \lambda_1 b) - \nabla F(a + \lambda_2 b), b\rangle \leq \frac{2L}{\kappa}\|\lambda_1 b - \lambda_2 b\|^\kappa$$

By taking the absolute value :

$$|g'(\lambda_1) - g'(\lambda_2)| \leq \frac{2L}{\kappa}\|\lambda_1 b - \lambda_2 b\|^\kappa = \frac{2L}{\kappa}\|b\|^\kappa |\lambda_1 - \lambda_2|^\kappa.$$

■

We notice that, since we use the binary search in the segment $[0, 1]$ (i.e. $g : [0, 1] \to \mathbb{R}$), F being (L, κ)-weakly smooth implies that g' is $\frac{2L}{\kappa}\|x_t - x_t^{ag}\|^\kappa$-Lipschitz.

Lemma 19 *Consider $a < b$ and let $g : \mathbb{R} \to \mathbb{R}$ have a continuous derivative. Then*

$$g(a), g'(b) > 0 \geq g(b) \implies \exists \lambda_\star \in [a, b], \quad g(\lambda_\star) \leq 0, g'(\lambda_\star) = 0$$

Proof Let us pick $\lambda_\star = \sup\{\lambda \in [a, b] \mid g'(\lambda) \leq 0\}$. That search set is not empty because due to Taylor-Lagrange, as g' is continuous, there exists $\lambda \in [a, b]$ such that $g'(\lambda) = \frac{g(b)-g(a)}{b-a} < 0$. By continuity of g and g', we conclude that $g(\lambda_\star) \leq 0, g'(\lambda_\star) = 0$. ■

Now we are ready to prove Theorem 7 on the complexity of the binary search.

Theorem 20 *For fixed $(C_t, \epsilon_t, x_t, x_t^{ag})$, the binary search Algorithm 2 finds a λ such that*

$$x_t^{md} = \lambda x_t^{ag} + (1 - \lambda)x_t$$
$$\langle \nabla F(x_t^{md}), x_t^{md} - x_t\rangle + C_t\Big(F(x_t^{md}) - F(x_t^{ag})\Big) \leq \epsilon_t \tag{19}$$

in $\mathcal{O}_\kappa\left(\max\{\log(C_t), \log\left(\frac{L\|x_t - x_t^{ag}\|^\kappa}{\epsilon_t}\right)\} + 1\right)$ iterations.

Proof Set

$$g(\lambda) := F(x_t^{md}) - F(x_t^{ag}) = F(\lambda x_t^{ag} + (1-\lambda)x_t) - F(x_t^{ag}).$$

We notice that $g(1) = 0$ and

$$g'(\lambda) = \langle \nabla F(x_t^{md}), x_t^{ag} - x_t \rangle$$
$$\langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle = \langle \nabla F(x_t^{md}), \lambda(x_t^{ag} - x_t) \rangle = \lambda g'(\lambda).$$

As a first step in our proof, we show that there indeed exists a λ satisfying (19), i.e., fulfilling

$$C_t g(\lambda) + \lambda g'(\lambda) \leq \epsilon_t.$$

If $\lambda = 0$ and $\lambda = 1$ fail to satisfy the inequality, we know that both :

$$g'(1) > \epsilon_t$$
$$C_t g(0) > \epsilon_t$$

From now, we will assume $g'(1), g(0) > 0$ for the rest of the proof. We set an index set $\Lambda_\star = \{\lambda \in [0,1] \mid g(\lambda) \leq 0, g'(\lambda) = 0\}$. We notice that

$$\lambda \in \Lambda_\star \implies C_t g(\lambda) + \lambda g'(\lambda) \leq 0.$$

We show that $\Lambda_\star$ is not an empty set. For that we will apply Lemma 19 with $g(0), g'(1) > 0 = g(1)$, there exists $\lambda_0 \in [0,1], g(\lambda_0) \leq 0, g'(\lambda_0) = 0$ and $\lambda_0 \in \Lambda_\star$.

Now for any $\lambda_\star \in \Lambda_\star$, since g' is α-Holder-smooth according to Lemma 18, for $\delta > 0$,

$$\forall \lambda \in [\lambda_\star - \delta, \lambda_\star] : \quad |g'(\lambda) - g'(\lambda_\star)| \leq \frac{2L}{\kappa} \|x_t - x_t^{ag}\|^\kappa \delta^\kappa$$
$$\forall \lambda \in [\lambda_\star - \delta, \lambda_\star] : \quad g(\lambda) - g(\lambda_\star) \leq \int_{\lambda_\star - \delta}^{\lambda_\star} |g'(t)| dt \leq \frac{2L}{\kappa} \|x_t - x_t^{ag}\|^\kappa \delta^{\kappa+1}$$

Therefore, for all $\lambda \in [\lambda_\star - \delta, \lambda_\star]$,

$$\begin{aligned}
& C_t g(\lambda) + \lambda g'(\lambda) \\
&= \underbrace{C_t g(\lambda_\star) + \lambda_\star g'(\lambda_\star)}_{\leq 0} + C_t(g(\lambda) - g(\lambda_\star)) + \lambda g'(\lambda) - \lambda_\star \underbrace{g'(\lambda_\star)}_{=0} \\
&\leq C_t \frac{2L}{\kappa} \|x_t - x_t^{ag}\|^\kappa \delta^{\kappa+1} + \lambda \frac{2L}{\kappa} \|x_t - x_t^{ag}\|^\kappa \delta^\kappa \\
&\leq \frac{2L}{\kappa} \|x_t - x_t^{ag}\|^\kappa \delta^\kappa (C_t \delta + 1).
\end{aligned}$$

In order to make the last expression smaller than ϵ_t, we consider $\delta = \min\{\frac{1}{C_t}, \left(\frac{\kappa \epsilon_t}{4L\|x_t - x_t^{ag}\|^\kappa}\right)^{1/\kappa}\}$. The last step is to notice that in Algorithm 2, our stop condition is satisfied when

$$\Lambda_\star \cap [a,b] \neq \emptyset, \delta > b - a,$$

since that implies $\exists \lambda_\star \in \Lambda_\star, a \in [\lambda_\star - \delta, \lambda_\star] \implies C_t g(a) + a g(a) \leq \epsilon_t$.

Now we verify that the condition $\Lambda_\star \in [a,b]$ holds for all the iterates. For that, using the previous Lemma 19, we only need to verify the conditions at the extremities of the segment. If we initially have :

$$g(a), g'(b) > 0 \geq g(b)$$

then while iterating with $\frac{a+b}{2}$, we know that :

$$g\Big(\frac{a+b}{2}\Big) > 0 \implies g\Big(\frac{a+b}{2}\Big), g'(b) > 0 \geq g(b) \implies \Lambda_\star \cap \Big[\frac{a+b}{2}, b\Big] \neq \emptyset$$
$$g'\Big(\frac{a+b}{2}\Big) > 0 \geq g\Big(\frac{a+b}{2}\Big) \implies g(a), g'\Big(\frac{a+b}{2}\Big) > 0 \geq g\Big(\frac{a+b}{2}\Big) \implies \Lambda_\star \cap \Big[a, \frac{a+b}{2}\Big] \neq \emptyset$$
$$0 \geq g\Big(\frac{a+b}{2}\Big), g'\Big(\frac{a+b}{2}\Big) \implies \frac{a+b}{2} \in \Lambda_\star$$

Once we prove that our iteration is valid, we know the interval starts as $[0,1]$ and we halve the size in each iteration until it reaches δ. So the maximum iteration count is :

$$\log(1/\delta) + 1 = \max\left\{\log(C_t), \log\left(\frac{4L\|x_t - x_t^{ag}\|^\kappa}{\kappa\epsilon_t}\right)^{1/\kappa}\right\} + 1$$

■

Appendix F. Lower bounds for binary search

We recall from Section 4 the sub-problem (6) of finding $\lambda \in [0,1]$ such that

$$Cg(\lambda) + \lambda g'(\lambda) \leq \epsilon \tag{20}$$

with $C, \epsilon > 0$ two constants and $g(1) = 0$. In this section, we show a lower bound on the number of iterations required to solve that problem in a first-order query model. We denote the class of weakly-smooth functions by

$$\mathcal{F}_{L_\star,\kappa} := \{g \in \mathcal{C}^1([0,1], \mathbb{R}) \mid g(1) = 0, \quad \forall a, b \in [0,1], \quad |g'(a) - g'(b)| \leq L_\star |a-b|^{\kappa-1}\},$$

and we denote the class of smooth functions by $\mathcal{F}_{L_\star} := \mathcal{F}_{L_\star,2}$. Since the domain is $[0,1]$, we know that $\mathcal{F}_{L_\star} \subset \mathcal{F}_{L_\star,\kappa}$ for all $1 < \kappa \leq 2$. The lower bound consists of building two smooth functions that agree with the observed function values and derivatives $(g(\lambda_t), g'(\lambda_t))$ for every λ_t evaluated, but these two functions have distinct intervals where the condition (20) is satisfied on the line. For simplicity, we refer to the condition (20) as the linear condition.

The following theorem gives the key condition on the interval $[a,b]$ to build counter examples. More precisely, we build g in the interval $[a,b]$ and we need to satisfy the interpolation condition $(g(a), g'(a)) = (A, 0)$ and $(g(b), g'(b)) = (B, B')$. C is the constant given in the problem (20) and $C_\star$ is related to C where we need $[a,b] \subset [1/C_\star, 1]$ so that our lower bound will work. For the upcoming theorem, we will consider the following notation

$$C_\star = 1 + \frac{1}{C}$$
$$\Phi(a,b,A,B,B') = \frac{56B'}{L_\star(b-a)} + \frac{32(A-B)}{L_\star(b-a)^2}. \tag{21}$$

Now, we can give the main theorem for lower bound.

Theorem 21 *If the following conditions are satisfied,*

$$\begin{cases} [a,b] \subset [1/C_\star, 1] \\ \Phi(a,b,A,B,B') \leq 1 \\ A \geq \frac{\epsilon}{C}, \qquad CB + \frac{B'}{C_\star} \geq \epsilon \\ B' \geq 0 \geq B \end{cases} \tag{22}$$

then there exists two functions $g_1, g_2 \in \mathcal{F}_{L_\star}$ such that :

$$\begin{aligned} g_1(a) = g_2(a) &= A \\ g_1(b) = g_2(b) &= B \\ g_1'(a) = g_2'(a) &= 0 \\ g_1'(b) = g_2'(b) &= B' \\ \forall \lambda \in \left[a, \frac{a+b}{2}\right], &\quad Cg_1(\lambda) + \lambda g_1'(\lambda) \geq \epsilon \\ \forall \lambda \in \left[\frac{a+b}{2}, b\right], &\quad Cg_2(\lambda) + \lambda g_2'(\lambda) \geq \epsilon \end{aligned}$$

We also notice that if the conditions are satisfied with 2ϵ instead of ϵ, we can have a strict guarantee on the last lines on ϵ, therefore the strictness of inequality is not important here. Theorem 21 claims that if the condition evaluation parameter Φ is smaller than 1, it is impossible to determine the location of the correct λ_t since there are still two functions in $\mathcal{F}_{L_\star}$ providing distinct intervals of solutions λ_t (by the continuity of g', a solution λ_t has to exist over the whole interval $[a,b]$ if the interpolation conditions on (a,b) are satisfied.) We provide the proof in the next section.

F.1. Proof with two separate constructions

In this section, we will show how to build the two smooth functions (g_1, g_2) in two different ways in theory. (In practice, one only need to build them at the end of all requests to show the lower bounds.) Both functions will interpolate (i.e. agree with) given first-order boundary conditions, but the first function g_1 has the property to not have any point satisfying (20) on the left side $\left[a, \frac{a+b}{2}\right]$ whereas the second function g_2 does not satisfy the condition (20) on the right side $\left[\frac{a+b}{2}, b\right]$. We use two separate constructions because (20) is not symmetric. The rough idea is that if the size of the interval $[a,b]$ is big enough, then the construction is possible while respecting the $L_\star$-smoothness condition.

Lemma 22 *We consider $[a,b] \subseteq [0,1]$. If the following conditions are satisfied :*

$$\begin{cases} b-a & \geq \frac{6B'}{L_\star} + \frac{16(A-B)}{L_\star(b-a)} \\ A & \geq \frac{\epsilon}{C} \\ B' & \geq 0 \end{cases} \tag{23}$$

then there exists an interpolating $L_\star$-smooth function g_1 such that :

$$\begin{aligned} g_1'(a) = 0, &\quad g_1(a) = A, \\ g_1'(b) = B', &\quad g_1(b) = B, \\ \forall x \in \left[a, a + \frac{b-a}{2}\right], &\quad Cg_1(\lambda) + \lambda g_1'(\lambda) \geq \epsilon \end{aligned}$$

Proof We will make a three-piece piece-wise linear interpolation of the derivative g':

$$g_1'\Big(a+\frac{b-a}{2}\Big)=0$$
$$g_1'\Big(a+\frac{3(b-a)}{4}\Big)=-W$$

with $W>0$ that we determine later. Now we will verify the linear condition (20) on different segments of $[a,b]$. For $\lambda\in\Big[a,a+\frac{b-a}{2}\Big]$, we have $g_1'(\lambda)=0,\quad g_1(\lambda)=A$. Therefore, in this interval, we can always assure that

$$Cg_1(\lambda)+\lambda g_1'(\lambda)\geq CA\geq\epsilon.$$

For $\lambda\in\Big[a+\frac{b-a}{2},a+\frac{3(b-a)}{4}\Big]$, we have the linear interpolation

$$g_1'(\lambda)=-\frac{4W}{b-a}\times\Big(\lambda-a-\frac{b-a}{2}\Big),$$

For $\lambda\in\Big[a+\frac{3(b-a)}{4},b\Big]$, we have the linear interpolation

$$g_1'(\lambda)=-W+\frac{4(B'+W)}{b-a}\times\Big(\lambda-a-\frac{3(b-a)}{4}\Big),$$

In the end we want to make sure that the integral of g' over $[a,b]$ is $B-A$:

$$g_1\Big(a+\frac{b-a}{2}\Big)+\int_{a+\frac{b-a}{2}}^{b}g_1'(\lambda)d\lambda=B.$$

We can develop the previous form, since g' is a linear interpolation, the integral is just the average of its two extreme points :

$$\begin{aligned}A-\frac{W(b-a)}{8}+\frac{b-a}{4}\times\frac{-W+B'}{2}&=B\\ \Leftrightarrow -W-W+B'&=\frac{8(B-A)}{b-a}\\ \Leftrightarrow W&=\frac{B'}{2}+\frac{4(A-B)}{b-a}\end{aligned}$$

Now that we find the value for W, we need to check the smoothness on the two last intervals.

$$\begin{cases}W & \leq L_\star\times\frac{b-a}{4}\\ B'+W & \leq L_\star\times\frac{b-a}{4}\end{cases}$$

Since we supposed that $B'\geq 0$, the second condition is stricter. We rewrite the second condition as :

$$\begin{aligned}&\frac{B'}{2}+\frac{4(A-B)}{b-a}+B'\leq L_\star\times\frac{b-a}{4}\\ \Leftrightarrow\ &\frac{6B'}{L_\star}+\frac{16(A-B)}{L_\star(b-a)}\leq(b-a)\end{aligned}$$

which is what we set out to show. ■

Now we will build the second interpolation function on the right side of the interval $\left[\frac{a+b}{2}, b\right]$. Unlike g_1, constant derivative is not working here, we show in the following lemma that linear derivative works if $[a, b]$ is big enough, but contained in $[1/C_\star, 1]$.

Lemma 23 *Let's consider $C_\star = 1 + \frac{1}{C}$. We consider $[a, b] \subset [1/C_\star, 1]$. If the following conditions are satisfied :*

$$\begin{cases} b - a & \geq \frac{56B'}{L_\star} + \frac{32(A-B)}{L_\star(b-a)} \\ B' & \geq 0 \geq B \\ CB + \frac{B'}{C_\star} & \geq \epsilon, \end{cases} \tag{24}$$

then there exists an $L_\star$-smooth function g_2 such that :

$$\begin{aligned} &g_2'(a) = 0, \quad g_2(a) = A \\ &g_2'(b) = B', \quad g_1(b) = B \\ &\forall \lambda \in \left[a + \frac{b-a}{2}, b\right], \quad Cg_2(\lambda) + \lambda g_2'(\lambda) \geq \epsilon \end{aligned}$$

Proof Now we start making linear interpolation for the derivatives with the middle point values fixed in the following way

$$\begin{aligned} g_2'\Big(a + \frac{b-a}{4}\Big) &= -W \\ g_2'\Big(a + \frac{b-a}{2}\Big) &= KB' \end{aligned}$$

with $V, W > 0$ two constants to be determine later. Now we will start with the right side as it is the most interesting one.

For $\lambda \in \left[a + \frac{(b-a)}{2}, b\right]$, we apply the linear interpolation :

$$\begin{aligned} g_2'(\lambda) &= B' + \frac{2(KB' - B')}{b-a} \times \Big(b - \lambda\Big) \\ &= B' \left(1 + \frac{2(K-1)}{b-a} \times \Big(b - \lambda\Big)\right) \end{aligned}$$

Since the interpolation of the derivative is linear, we know that :

$$\begin{aligned} g_2(b) - g_2(\lambda) &= \int_\lambda^b g'(s)ds \\ &= \Big(\frac{g_2'(b) + g_2'(\lambda)}{2}\Big)(b - \lambda) \\ g_2(\lambda) &= B - B' \left(1 + \frac{(K-1)(b-\lambda)}{b-a}\right)(b - \lambda) \end{aligned}$$

Since $a \geq 1/3$ and $g_2' \geq 0$ on this interval, we know that:

$$\begin{aligned}
&Cg_2(\lambda) + xg_2'(\lambda) \\
\geq& Cg_2(\lambda) + \frac{g_2'(\lambda)}{C_\star} \\
\geq& CB - CB'\left(1 + \frac{(K-1)(b-\lambda)}{b-a}\right)(b-\lambda) + \frac{B'}{C_\star}\left(1 + \frac{2(K-1)}{(b-a)} \times (b-\lambda)\right) \\
\geq& \Big(CB + \frac{B'}{C_\star}\Big) + \left(\frac{2(K-1)}{C_\star(b-a)} - C\Big(1 + \frac{(K-1)(b-\lambda)}{b-a}\Big)\right) \times B'(b-\lambda)
\end{aligned}$$

We only need to choose K big enough so $\frac{2(K-1)}{C_\star(b-a)} - C\Big(1 + \frac{(K-1)(b-\lambda)}{b-a}\Big)$ is always positive, we notice that :

$$\begin{aligned}
\frac{b-\lambda}{b-a} &\leq \frac{1}{2} \\
b - a &\leq 1 - \frac{1}{C_\star} = \frac{C_\star - 1}{C_\star} \\
\implies \frac{1}{C_\star(b-a)} &\geq \frac{1}{C_\star}\frac{C_\star}{C_\star - 1} = \frac{1}{C_\star - 1}
\end{aligned}$$

then

$$\begin{aligned}
&\frac{2(K-1)}{C_\star(b-a)} - C\Big(1 + \frac{(K-1)(b-\lambda)}{b-a}\Big) \\
\geq& \frac{2(K-1)}{C_\star - 1} - C\Big(1 + \frac{K-1}{2}\Big) \\
=& \frac{2(K-1)}{C_\star - 1} - \frac{C(K+1)}{2} \\
=& \frac{4}{C_\star - 1} - 2C \geq 0
\end{aligned}$$

by picking $K = 3$ in the last line and as we recall that $C_\star = 1 + \frac{1}{C}$.

For $x \in \Big[a, a + \frac{(b-a)}{2}\Big]$, we need to assure the integral of g' over $[a, b]$

$$\begin{aligned}
\int_a^{a+\frac{b-a}{2}} g_2'(\lambda)d\lambda = g_2\Big(\frac{a+b}{2}\Big) - g_2(a) &= B - B'\left(1 + \frac{K-1}{2}\right)\frac{b-a}{2} - A \\
&= B - B'(b-a) - A
\end{aligned}$$

We are making an linear interpolation of g_2' and

$$g_2'(a) = 0, \quad g_2'\Big(a + \frac{b-a}{4}\Big) = -W, \quad g_2'\Big(a + \frac{b-a}{2}\Big) = 3B',$$

then,

$$\begin{aligned}
\frac{0 + 3B' - 2W}{4} \times \frac{b-a}{2} &= B - A - B'(b-a) \\
\Leftrightarrow 3B' - 2W + 8B' &= \frac{8(B-A)}{b-a} \\
W &= 11B' + \frac{8(A-B)}{b-a}
\end{aligned}$$

The remaining task is to check the $L_\star$-smoothness over the three segments :

$$\begin{cases} W & \leq L_\star \times \frac{b-a}{4} \\ W + 3B' & \leq L_\star \times \frac{b-a}{4} \\ 2B' & \leq L_\star \times \frac{b-a}{2} \end{cases}$$

The second one is the most restrictive one. It is equiavalent to :

$$\begin{aligned} 11B' + \frac{8(A-B)}{b-a} + 3B' &\leq L_\star \times \frac{b-a}{4} \\ \Leftrightarrow \frac{56B'}{L_\star} + \frac{32(A-B)}{L_\star(b-a)} &\leq b-a \end{aligned}$$

■

F.2. Induced Lower bound theory

Now combining Lemma 22 with Lemma 23, we will design a lower bound adversary $\mathcal{A}$ that limits the growth of Φ from (21) at each evaluation.

If $\lambda \in \left[a, a + \frac{b-a}{2}\right]$, $\mathcal{A}$ returns :

$$g'(\lambda) = \epsilon, \quad g(\lambda) = A$$

If $\lambda \in \left[a + \frac{b-a}{2}, b\right]$, $\mathcal{A}$ returns :

$$g'(\lambda) = B'\left(1 + \frac{4(b-\lambda)}{b-a}\right), \quad g(\lambda) = B - B'\left(1 + \frac{2(b-\lambda)}{b-a}\right)(b-\lambda)$$

Theorem 24 *We consider the same condition as in Theorem 21. After requesting (g, g') at λ, the algorithm $\mathcal{A}$ returns $(g(\lambda), g'(\lambda))$. We know that value of Φ on $[a', b'] = [a, \lambda]$ or $[a', b'] = [\lambda, b]$ will verify that :*

$$\begin{cases} \Phi\Big(a', b', g(a'), g(b'), g'(b')\Big) \leq 5\Phi(a, b, A, B, B') \\ g(a') \geq \frac{\epsilon}{C}, \quad g'(b') \geq 0 \geq g(b') \\ Cg(b') + \frac{g'(b')}{C_\star} \geq \epsilon, \end{cases}$$

Proof If $\lambda \in \left[a, a + \frac{b-a}{2}\right]$, the new segment is on the right side of λ, $[a', b'] = [\lambda, b]$. Since $b - a \leq 2(b - \lambda)$, we know that

$$\begin{aligned} \Phi\Big(a', b', g(a'), g(b'), g'(b')\Big) &= \frac{56g'(b')}{L_\star(b-\lambda)} + \frac{32(g(\lambda) - g(b))}{L_\star(b-\lambda)^2} \\ &= \frac{56B'}{L_\star(b-\lambda)} + \frac{32(A-B)}{L_\star(b-\lambda)^2} \\ &\leq 2 \times \frac{56B'}{L_\star(b-a)} + 4 \times \frac{32(A-B)}{L_\star(b-a)^2} \leq 4\Phi(a, b, A, B, B') \end{aligned}$$

If $\lambda \in \left[a + \frac{b-a}{2}, b\right]$, the new segment is on the left side of λ, $[a', b'] = [a, \lambda]$. We recall that

$$g'(\lambda) = B'\left(1 + \frac{4(b-\lambda)}{b-a}\right), \quad g(\lambda) = B - B'\left(1 + \frac{2(b-\lambda)}{b-a}\right)(b-\lambda)$$

then

$$\begin{aligned}
&\Phi\Big(a', b', g(a'), g(b'), g'(b')\Big)\\
=&\frac{56g'(b')}{L_\star(\lambda-a)} + \frac{32(g(\lambda)-g(b))}{L_\star(\lambda-a)^2}\\
=&\frac{56g'(\lambda)}{L_\star(\lambda-a)} + \frac{32(A-g(\lambda))}{L_\star(\lambda-a)^2}\\
=&\frac{56B'}{L_\star(\lambda-a)} + \frac{56B'}{L_\star(\lambda-a)} \times \frac{4(b-\lambda)}{b-a} + \frac{32(A-B)}{L_\star(\lambda-a)^2} + \frac{32B'}{L_\star(\lambda-a)^2} \times \frac{2(b-\lambda)^2}{b-a}\\
=&\frac{56B'}{L_\star(\lambda-a)} + \frac{56B'}{L_\star(\lambda-a)} \times \frac{4(b-\lambda)}{b-a} + \frac{32B'}{L_\star(b-a)} \times \frac{2(b-\lambda)^2}{(\lambda-a)^2} + \frac{32(A-B)}{L_\star(\lambda-a)^2}\\
\leq&\frac{56B'}{L_\star(\lambda-a)} + \frac{56B'}{L_\star(\lambda-a)} \times 2 + \frac{32B'}{L_\star(b-a)} \times \frac{1}{2} + \frac{32(A-B)}{L_\star(\lambda-a)^2}
\end{aligned}$$

with the last inequality using that $2(b-\lambda) \leq b-a$, then

$$\begin{aligned}
&\Phi\Big(a', b', g(a'), g(b'), g'(b')\Big)\\
\leq&\frac{5}{2} \times \frac{56B'}{L_\star(\lambda-a)} + \frac{32(A-B)}{L_\star(\lambda-a)^2}\\
\leq&5 \times \frac{56B'}{L_\star(b-a)} + 4 \times \frac{32(A-B)}{L_\star(b-a)^2} = 5\Phi(a, b, A, B, B')
\end{aligned}$$

Combining both, we also guarantee that

$$\Phi(a', b', g(a'), g(b'), g'(b')) \leq 5\Phi(a, b, A, B, B')$$

■

In the previous theorem, we showed that the condition evaluation function Φ grows at an exponential speed at most through each evaluation and we know that $\Phi \leq 1$ means that we can always find two smooth functions which have distinct intervals that satisfy (20). Combining both, the next theorem gives an explicit formulation for the logarithmic complexity

Theorem 25 *Let's fix parameters $C, \epsilon, L_\star > 0$. Consider a sequence of N points $(\lambda_1, \dots, \lambda_N)$ where $(g(\lambda_i), g'(\lambda_i))$ are evaluated. If the number of iterations is insufficient*

$$N < \log(5)\Big(\log\frac{L_\star}{\epsilon} + \log\frac{C}{(C+1)^2} - \log(88)\Big)$$

Then there exists two $L_\star$-smooth functions (g_1, g_2) and two sets I_1, I_2 such that

$$\begin{aligned}
\forall 1 \leq i \leq N, \quad & g_1(\lambda_i) = g_2(\lambda_i) = g(\lambda_i) \\
\forall 1 \leq i \leq N, \quad & g_1'(\lambda_i) = g_2'(\lambda_i) = g'(\lambda_i) \\
\forall \lambda \in I_1, \quad & C g_1(\lambda) + \lambda g_1'(\lambda) \geq \epsilon \\
\forall \lambda \in I_2, \quad & C g_2(\lambda) + \lambda g_2'(\lambda) \geq \epsilon \\
& I_1 \cup I_2 = [0, 1]
\end{aligned}$$

Proof The lower depends on the initial value of Φ. We can consider $[a, b] = [\frac{1}{C_\star}, 1]$ with $A = \frac{\epsilon}{C}$, $B = 0$ and $B' = C_\star \epsilon$. We recall that $C_\star = \frac{C+1}{C}$ $(1 - \frac{1}{C_\star} = \frac{1}{C+1})$ and we notice that

$$\begin{aligned}
\Phi\left(\frac{1}{C_\star}, 1, \frac{\epsilon}{C}, 0, C_\star \epsilon\right) &= \frac{56 C_\star \epsilon (C+1)}{L_\star} + \frac{32 \epsilon (C+1)^2}{C L_\star} \\
&= \frac{56 \epsilon (C+1)^2}{C L_\star} + \frac{32 \epsilon (C+1)^2}{C L_\star} = \frac{88 \epsilon (C+1)^2}{C L_\star}
\end{aligned}$$

Then the minimum iteration is

$$\log_5\left(\frac{C L_\star}{88 \epsilon (C+1)^2}\right) = \log(5)\Big(\log \frac{L_\star}{\epsilon} + \log \frac{C}{(C+1)^2}\Big) + \mathcal{O}(1)$$

■

Appendix G. Lower bounds in p-norm

For the lower bound, we use the results from to Diakonikolas and Guzmán (2024) where the authors consider convex functions that are weakly smooth in a non-Euclidean norm.

Theorem 26 ((Diakonikolas and Guzmán, 2024)) *Let $1 \leq p \leq \infty$, and consider the problem class of unconstrained minimization with objectives in the class $\mathcal{F}_{\mathbb{R}^d, \|\cdot\|_p}(\kappa, L)$, whose minima are attained in $\mathcal{B}_{\|\cdot\|_p}(0, R)$. Then, the complexity of achieving additive optimality gap ϵ, for any local oracle, is bounded below by:*

$$\begin{cases}
\Omega\left(\left(\frac{L R^\kappa}{\epsilon [\ln d]^{\kappa-1}}\right)^{\frac{2}{3\kappa-2}}\right), & \text{if } 1 \leq p < 2; \\
\Omega\left(\left(\frac{L R^\kappa}{\epsilon \min\{p, \ln d\}^{\kappa-1}}\right)^{\frac{p}{\kappa p + \kappa - p}}\right), & \text{if } 2 \leq p < \infty; \text{ and,}
\end{cases}$$

The dimension d for the lower bound to hold must be at least as large as the lower bound itself.

We would like to point out that the previous lower bounds use convex functions whereas the upper bounds provided in our work focus on the star-convex function class. Therefore, it is surprising that our algorithms performs nearly optimally on p-norms up to a factor depending only on τ.

Appendix H. Bregman divergence and radius of domain

In this section we relate the Bregman divergence generated by the q-th power of the p-norm to the q-th power of the p-norms of the arguments.

Lemma 27 *Fix $p > 1$, $q > 1$ and define $\psi(x) := \frac{1}{q}\|x\|_p^q$. Then*

$$D_\psi(x, y) \;\le\; 2\max\{\,\|x\|_p^q\,, \|y\|_p^q\,\}.$$

Proof By definition, we have

$$D_\psi(x, y) \;=\; \frac{1}{q}\|x\|_p^q - \frac{1}{q}\|y\|_p^q - \|y\|_p^{q-p}\sum_i (x_i - y_i)|y_i|^{p-2}y_i,$$

which we may reorganize to

$$\le\; \frac{1}{q}\|x\|_p^q + \frac{q-1}{q}\|y\|_p^q + \|y\|_p^{q-p}\sum_i x_i|y_i|^{p-1}.$$

Seeing that sum as an inner product, by Hölder with $1 = \frac{1}{p} + \frac{1}{\frac{p}{p-1}}$, this is bounded above by

$$\le\; \frac{1}{q}\|x\|_p^q + \frac{q-1}{q}\|y\|_p^q + \|x\|_p\|y\|_p^{q-1}$$

and further by convexity of the exponential (i.e. $a^\theta b^{1-\theta} = e^{\theta\ln a + (1-\theta)\ln b} \le \theta a + (1-\theta)b$) by

$$\le\; 2\left(\frac{1}{q}\|x\|_p^q + \frac{q-1}{q}\|y\|_p^q\right).$$

The result follows by observing that a convex combination is bounded by the maximum. ■

Appendix I. Application for l1 norm

For the ℓ_1 norm, we can apply our algorithm for ℓ_q norm with $q = 1 + s$ and use the equivalence of norms :

$$D_F(x,y) \le \frac{L}{\kappa}\|x-y\|_1^\kappa \le \frac{Ld^{\kappa(1-\frac{1}{q})}}{\kappa}\|x-y\|_q^\kappa = \frac{Ld^{\frac{\kappa s}{s+1}}}{\kappa}\|x-y\|_q^\kappa$$

Since F is $(Ld^{\frac{\kappa s}{s+1}}, \kappa)$-weakly-smooth with respect to the q-norm, we can pick $\psi(x) = \frac{1}{2}\|x\|_q^2$ as the distance generating function; it is strongly convex with respect to the q-norm.

Now we have a weakly smooth function and a strongly convex distance generating function with respect to q-norm, our algorithms leads to the following precision

$$\mathcal{O}_\kappa\left(d^{\frac{2\kappa s}{(s+1)(3\kappa-2)}}\left(\frac{L\tau^2R^\kappa}{\epsilon}\right)^{\frac{2}{3\kappa-2}}\log^2\left(\frac{L\tau R}{\epsilon}\right)\right)$$

with $s > 0$ a parameter that can be chosen arbitrary small. Therefore, the dimension cost can be controlled asymptotically when $d \to \infty, \epsilon \to 0$.

Appendix J. Smooth case where q=2

For the completeness of the theory, in the special case where $q = \kappa = 2$, we present a variation of our Theorem 5 here. We start by recalling the definitions of the standard smoothness and strong convexity.

Definition 28 (Smoothness) *A continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be L-smooth with respect to norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^d$*

$$|D_F(x, y)| \leq \frac{L}{2}\|x - y\|^2.$$

Definition 29 (Strong convexity) *A continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be μ-strongly convex with respect to norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^d$*

$$D_F(x, y) \geq \frac{\mu}{2}\|x - y\|^2.$$

Theorem 30 *In the setting of Assumption 1, Algorithm 1 with the tuning below gives for all $t \geq 1$*

$$A_t\Big(F(x_{t+1}^{ag}) - F_\star\Big) \leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t\epsilon_t + A_{t-1}\Big(F(x_t^{ag}) - F_\star\Big)$$

where $\mathcal{O}_{q,\kappa}$ omits constants depending only on (q, κ). This is achieved for any $\alpha > 0$ by

$$\alpha_t := \alpha, \qquad \eta_t := \frac{\alpha t}{2\tau}, \qquad \epsilon_t := \frac{1}{t\eta_t}, \qquad A_t := \frac{\eta_t^2}{\alpha_t}. \tag{25}$$

Proof Compared to the case where $q > \kappa$, the step is the same until equation (11), then the smooth analysis changes.

$$\begin{aligned}
&\frac{\eta_t}{\tau}\Big(F(x_t^{md}) - F_\star\Big) \\
&\leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t\langle\nabla F(x_t^{md}), x_t^{md} - x_t\rangle \\
&\quad + \eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle - \frac{\mu}{2}\|x_{t+1} - x_t\|^2.
\end{aligned} \tag{26}$$

For the smooth analysis part, we still obtain :

$$\begin{aligned}
\alpha_t\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) &\geq \mu\|x_t^{md} - x_{t+1}^{ag}\|^2 - \alpha_t D_F(x_{t+1}^{ag}, x_t^{md}) \\
&\geq \mu\|x_t^{md} - x_{t+1}^{ag}\|^2 - \frac{L\alpha_t}{2}\|x_t^{md} - x_{t+1}^{ag}\|^2,
\end{aligned} \tag{27}$$

We will assume $\frac{2\mu}{L} > \alpha_t$. Similarly to the case $q > \kappa$, the smoothness of F combining with the choice of x_{t+1}^{ag} leads to

$$\begin{aligned}
&\eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle - \frac{\mu}{2}\|x_{t+1} - x_t\|^2 \\
&\leq \frac{\mu\eta_t^2}{\alpha_t^2}\frac{\|x_{t+1}^{ag} - x_t^{md}\|^2}{2} \\
&\leq \frac{\mu\eta_t^2}{2\alpha_t^2}\left(\frac{2\alpha_t}{2\mu - L\alpha_t}\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big)\right) \\
&= \frac{\mu\eta_t^2}{\alpha_t(2\mu - L\alpha_t)}\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big).
\end{aligned} \tag{28}$$

Combining with previous equations (26) (28), we know that

$$\begin{aligned}\frac{\eta_t}{\tau}\Big(F(x_t^{md})-F_\star\Big) \leq D_\psi(x_\star,x_t)-D_\psi(x_\star,x_{t+1})+\eta_t\langle\nabla F(x_t^{md}),x_t^{md}-x_t\rangle \\ +\frac{\mu\eta_t^2}{\alpha_t(2\mu-L\alpha_t)}\Big(F(x_t^{md})-F(x_{t+1}^{ag})\Big).\end{aligned} \tag{29}$$

which can be arranged into :

$$\begin{aligned}&\frac{\mu\eta_t^2}{\alpha_t(2\mu-L\alpha_t)}\Big(F(x_{t+1}^{ag})-F_\star\Big)\\ \leq& D_\psi(x_\star,x_t)-D_\psi(x_\star,x_{t+1})+\eta_t\langle\nabla F(x_t^{md}),x_t^{md}-x_t\rangle\\ &+\Big(\frac{\mu\eta_t^2}{\alpha_t(2\mu-L\alpha_t)}-\frac{\eta_t}{\tau}\Big)\Big(F(x_t^{md})-F(x_t^{ag})\Big)\\ &+\Big(\frac{\mu\eta_t^2}{\alpha_t(2\mu-L\alpha_t)}-\frac{\eta_t}{\tau}\Big)\Big(F(x_t^{ag})-F_\star\Big).\end{aligned} \tag{30}$$

with binary search

$$\eta_t\langle\nabla F(x_t^{md}),x_t^{md}-x_t\rangle+\frac{\mu\eta_t^2}{\alpha_t(2\mu-L\alpha_t)}\Big(F(x_t^{md})-F(x_t^{ag})\Big)\leq\eta_t\epsilon_t$$

We write $A_t:=\frac{\eta_t^2}{\alpha_t}$, then

$$A_t\Big(F(x_{t+1}^{ag})-F_\star\Big)\leq D_\psi(x_\star,x_t)-D_\psi(x_\star,x_{t+1})+\eta_t\epsilon_t+A_{t-1}\Big(F(x_t^{ag})-F_\star\Big). \tag{31}$$

■

However, when it comes to the analysis of the upper bound of the iterates, an additional assumption needs to be considered. If the following condition is not satisfied, we cannot theoretically guarantee the polynomial growth of our iterates.

Assumption 2 *There exists $C\geq 1$, such that for all $x\in\mathbb{R}^d$,*

$$\|\nabla\psi(x)\|_*\leq C\|x\|$$

We note that this assumption is verified for p-norms with $1<p<2$ where we consider $\psi(x):=\frac{\|x\|_p^2}{2}$. There the assumption is equivalent to $\|\phi(x)\|_*\leq C\|x\|$ with $\phi(x):=\nabla\Big(\frac{\|x\|^2}{2}\Big)$. That condition is satisfied for $C=1$. Now, we can provide the upper bound for the bounded iterates.

Theorem 31 (case $q=\kappa=2$) *We suppose that :*

$$\forall x\in\mathbb{R}^d,\qquad \|\nabla\psi(x)\|_*\leq C\|x\|$$

Let's assume that $D_\psi(x_\star,x_1)\leq B$. We suppose that we are running Algorithm 1 with exponents $n_1\geq 0, n_2\geq 1$, such that for all $s\geq 1$

$$\max\left(\sum_{t=1}^s\eta_t\epsilon_t,\frac{\alpha_t}{\eta_t},L\eta_s,\frac{L\alpha_s}{2}\right)\leq K^{n_1}s^{n_2}$$

and $K^{n_1}>B$ that might also depend on (C,L,τ,μ). Then, there exists an upper bound for $\|x_t-x_t^{ag}\|$ which is polynomial in t.

Proof We use the result from Theorem 5, by telescopic sum, we know that :

$$D_\psi(x_\star, x_{\tau+1}) + A_\tau\Big(F(x^{ag}_{\tau+1}) - F(x_\star)\Big) \leq D_\psi(x_\star, x_1) + A_1\Big(F(x_1) - F(x_\star)\Big) + \sum_{t=1}^{\tau}(B_t + \eta_t\epsilon_t)$$

We recall that from the smoothness of F:

$$\begin{aligned}\|\nabla F(x^{md}_t)\|_* &\leq \frac{L}{2}\|x^{md}_t - x_\star\| \\ F(x_1) - F(x_\star) \leq D^F(x_1, x_\star) &\leq \frac{L}{2}\|x_1 - x_\star\|^2 \leq \frac{qL}{2}B\end{aligned}$$

which implies that

$$\begin{aligned}\mu\|x_{\tau+1} - x_\star\|^2 &\leq B + LB + K^{n_1}\tau^{n_2} \\ \|x_{\tau+1} - x_\star\| &\leq \Big(\frac{B}{\mu} + \frac{LB}{\mu} + \frac{K^{n_1}}{\mu}\tau^{n_2}\Big)^{\frac{1}{2}} \\ &\leq \Big(\frac{B}{\mu} + \frac{LB}{\mu}\Big)^{\frac{1}{2}} + \frac{K^{\frac{n_1}{2}}}{\sqrt{\mu}}\tau^{\frac{n_2}{2}}\end{aligned}$$

Since our setting is unconstrained, the (sub-)gradients have to cancel :

$$\begin{aligned}\eta_t\nabla F(x^{md}_t) + \nabla\psi(x_{t+1}) - \nabla\psi(x_t) &= 0 \\ \alpha_t\nabla F(x^{md}_t) + \phi(x^{ag}_{t+1} - x^{md}_t) &= 0\end{aligned}$$

with $\phi(x) := \nabla\Big(\frac{\|x\|^2}{2}\Big)$. With some manipulation

$$\begin{aligned}\eta_t\|\nabla F(x^{md}_t)\|_* &\leq \|\nabla\psi(x_{t+1})\|_* + \|\nabla\psi(x_t)\|_* \leq C\Big(\|x_{t+1}\| + \|x_t\|\Big) \\ &\leq C\Big(\|x_{t+1} - x_\star\| + \|x_t - x_\star\| + 2\|x_\star\|\Big) \\ \alpha_t\|\nabla F(x^{md}_t)\|_* &= \|\phi(x^{ag}_{t+1} - x^{md}_t)\|_* = \|x^{ag}_{t+1} - x^{md}_t\|\end{aligned}$$

We use the definition of ϕ, we combine the previous step with Cauchy–Schwarz inequality

$$\begin{aligned}\|x^{ag}_{t+1} - x^{md}_t\|^2 &\leq \langle\phi(x^{ag}_{t+1} - x^{md}_t), x^{ag}_{t+1} - x^{md}_t\rangle \\ &\leq \alpha_t\|\nabla F(x^{md}_t)\|_* \times \|x^{ag}_{t+1} - x^{md}_t\| \\ \|x^{ag}_{t+1} - x^{md}_t\| &\leq \alpha_t\|\nabla F(x^{md}_t)\|_*\end{aligned}$$

From the triangle inequality, we obtain :

$$\begin{aligned}\|x^{ag}_{t+1} - x_\star\| &\leq \|x^{ag}_{t+1} - x^{md}_t\| + \|x^{md}_t - x_\star\| \\ &\leq \alpha_t\|\nabla F(x^{md}_t)\|_* + \lambda_t\|x^{ag}_t - x_\star\| + (1-\lambda_t)\|x_t - x_\star\| \\ &\leq \frac{C\alpha_t}{\eta_t}\Big(\|x_{t+1} - x_\star\| + \|x_t - x_\star\| + 2\|x_\star\|\Big) + \|x_t - x_\star\| + \|x^{ag}_t - x_\star\|\end{aligned}$$

By induction,

$$\|x_{\tau+1}^{ag} - x_\star\| \leq \sum_{t=0}^{\tau} \left(\frac{C\alpha_{t+1}}{\eta_{t+1}} + \frac{C\alpha_t}{\eta_t} + 1 \right) \|x_{t+1} - x_\star\| + C\|x_\star\| \sum_{t=1}^{\tau} \frac{\alpha_t}{\eta_t}$$
$$\leq CK^{n_1} \sum_{t=0}^{\tau} \left((t+1)^{n_2} + t^{n_2} + 1 \right) \|x_{t+1} - x_\star\| + CK^{n_1} \|x_\star\| \tau^{n_2}$$

As we recall that

$$\|x_{t+1} - x_\star\| \leq \left(\frac{B}{\mu} + \frac{LB}{\mu} \right)^{\frac{1}{2}} + \frac{K^{\frac{n_1}{2}}}{\sqrt{\mu}} t^{\frac{n_2}{2}},$$

we know that $\|x_{t+1} - x_\star\|$ has an upper bound polynomial in t. Therefore, the distance $\|x_t - x_t^{ag}\|$ is polynomial in t as well. ■

Combining Theorem 30 with Theorem 31, we obtain the final convergence rate :

Corollary 32 (case $q = \kappa = 2$) *We suppose that :*

$$\forall x \in \mathbb{R}^d, \qquad \|\nabla \psi(x)\|_* \leq C\|x\|$$

If a bound $\frac{1}{\mu} D_\psi(x_\star, x_1) \leq B$ is available, then the tuning of Theorem 30 with $\alpha = \frac{\mu\sqrt{B}}{L}$ guarantees $F(x_{T+1}^{ag}) - F_\star = \mathcal{O}\left(\frac{L\tau^2\sqrt{B}}{T^2} \right)$. The algorithm's oracle usage over T iterations is upper-bounded by $\mathcal{O}(T \log(LB\tau T))$, where this bound represents the maximum number of times the oracle is called during the entire execution process.

Appendix K. Discussion on the extension to the constrained case

In the following we highlight the main challenges that arise when extending our result to the constrained case. Specifically, consider a convex domain $\mathcal{X} \subseteq \mathbb{R}^d$. The goal is to solve the constrained problem

$$\min_{x \in \mathcal{X}} F(x)$$

where F is assumed to be smooth and star-convex. For simplicity, set $q = \kappa = 2$. Our Algorithm 1 can be adapted to this constrained setting as follows:

$$x_t^{md} = \lambda_t x_t^{ag} + (1 - \lambda_t) x_t$$
$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{ \eta_t \langle \nabla F(x_t^{md}), x \rangle + D_\psi(x, x_t) \}$$
$$x_{t+1}^{ag} = \arg\min_{x \in I_t \cap \mathcal{X}} \{ \alpha_t \langle \nabla F(x_t^{md}), x \rangle + \frac{\mu}{2} \|x - x_t^{md}\|^2 \},$$

where $I_t := \{ x_t^{md} + \gamma(x_{t+1} - x_t), \gamma \in \mathbb{R} \}$ is the line passing through x_t^{md} and x_{t+1}. The original convergence argument until (11) applies for any x_{t+1}^{ag}, so we get

$$\frac{\eta_t}{\tau} \left(F(x_t^{md}) - F_\star \right)$$
$$\leq D_\psi(x_\star, x_t) - D_\psi(x_\star, x_{t+1}) + \eta_t \langle \nabla F(x_t^{md}), x_t^{md} - x_t \rangle \tag{32}$$
$$+ \eta_t \langle \nabla F(x_t^{md}), x_t - x_{t+1} \rangle - \frac{\mu}{2} \|x_{t+1} - x_t\|^2.$$

The main difficulty lies in bounding the term:

$$\eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle - \frac{\mu}{2}\|x_{t+1} - x_t\|^2 \leq \frac{\mu\eta_t^2}{\alpha_t(2\mu - L\alpha_t)}\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big)$$

using the new definition of x_{t+1}^{ag} from above. Interestingly, our original choice of x_{t+1}^{ag} can be seen to maximize a lower bound on $F(x_t^{md}) - F(x_{t+1}^{ag})$ by observing that

$$\begin{aligned} F(x_t^{md}) - F(x_{t+1}^{ag}) &= \langle\nabla F(x_t^{md}), x_t^{md} - x_{t+1}^{ag}\rangle - D_F(x_{t+1}^{ag}, x_t^{md}) \\ &\geq \langle\nabla F(x_t^{md}), x_t^{md} - x_{t+1}^{ag}\rangle - \frac{L}{2}\|x_{t+1}^{ag} - x_t^{md}\|^2. \end{aligned}$$

However, our proof strategy from the unconstrained case only applies if x_{t+1}^{ag} lies in the interior of $\mathcal{X}$. To see what fails, our proof in Section C.2 essentially relies on the unconstrained x_{t+1}^{ag} zeroing the derivative of its defining objective in the direction $x_{t+1} - x_t$. For the constrained x_{t+1}^{ag}, there is no useful relation between the derivative and that vector. We may still make some progress as follows. Let us denote the optimizer over the line by

$$x_{t+1}^{ag} = x_t^{md} + \beta_t(x_{t+1} - x_t).$$

The difficult case is when $\beta_t > 0$. In this case, one can still obtain :

$$\begin{aligned} \eta_t\langle\nabla F(x_t^{md}), x_t - x_{t+1}\rangle &= \frac{\eta_t}{\beta_t}\langle\nabla F(x_t^{md}), x_t^{md} - x_{t+1}^{ag}\rangle \\ &= \frac{\eta_t}{\beta_t}\Big(F(x_t^{md}) - F(x_{t+1}^{ag}) + D_F(x_{t+1}^{ag}, x_t^{md})\Big) \\ &\leq \frac{\eta_t}{\beta_t}\Big(F(x_t^{md}) - F(x_{t+1}^{ag})\Big) + \frac{L\eta_t\beta_t}{2}\|x_{t+1} - x_t\|^2, \end{aligned}$$

but we can only guarantee the accelerated convergence rate if $\beta_t \geq \frac{\alpha_t(2\mu - L\alpha_t)}{\mu\eta_t}$. This condition may not be satisfied in the constrained setting since β_t can be arbitrarily close to zero, especially when x_t^{md} lies near the boundary of $\mathcal{X}$. This is the key reason our analysis does not straightforwardly generalize to the constrained case.

Appendix L. Star-convexity and quasi-convexity

Following the definition of quasi-convexity from Boyd and Vandenberghe (2004), quasi-convexity and star-convexity are not hierarchically related. Quasi-convexity requires all sublevel sets to be convex (i.e., a condition on all pairs (x, y)), whereas τ-star-convexity only constrains the function along rays from the minimizer x^*. These impose different geometric requirements, and neither assumption is strictly stronger in general. For example, consider

$$g(x, y) = (x^2 + y^2)\, q(x, y)^2(1 - q(x, y))^2, \qquad q(x, y) = \frac{\arctan(|y|/|x|)}{\pi/2}.$$

The function is minimized at the origin and is a symmetric convex quadratic (with quadratic coefficient in $[0, 1/8]$) along every line through the origin, so g is star convex with $\tau = 1$. However, its level sets consist of all points within some distance to the coordinate axes (i.e. an infinite cross at the origin), and as such g is not quasi-convex.

Appendix M. Non-Accelerated Mirror Descent for Benchmarking

As a benchmark for minimizing τ-star-convex functions without acceleration, we employ the standard mirror descent algorithm. The update rule at iteration t is given by:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\eta \langle \nabla F(x_t), x \rangle + D_\psi(x, x_t)\}.$$

Convergence Result: Applying an analysis analogous to the accelerated case, with step size $\eta = \frac{D^{q-\kappa}}{L}$ and selecting the output as $x_{\text{out}} = \arg\min_{1 \le t \le T} F(x_{t+1})$, the non-accelerated algorithm achieves the following guarantee after T iterations:

$$F(x_{\text{out}}) - F_\star = \mathcal{O}\left(\frac{LD^\kappa}{\tau T^{\frac{\kappa}{q}}}\right),$$

where the parameters satisfy $1 < \kappa \le 2$ and $q \ge 2$.

Appendix N. Handling Composite Functions with Non-Euclidean Regularization

Our framework is particularly effective for composite objectives where one component is star-convex and the other is smooth with respect to a non-Euclidean norm. We demonstrate this with a concrete example.

Constructing the Example: Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(x) := g(x) + L_2\, h_p(x - \mathbf{1}_d),$$

where:

- $g(x)$ is the lower-bound function from Hinder et al. (2020), which is τ-star-convex.
- $h_p(\cdot)$ is the lower-bound function from Guzmán and Nemirovski (2015) for $1 < p \le 2$.

Properties of the Composite Function: The constructed function f possesses the following key characteristics:

1. **Smoothness:** f is $\mathcal{O}(L_2 + \Theta(L_2))$-smooth with respect to the ℓ_p-norm.
2. **Star-Convexity:** f is τ-star-convex.
3. **Non-Convexity:** For sufficiently small L_2, the function f is non-convex.
4. **Dimension Dependency:** f is not smooth with respect to the Euclidean norm without incurring a dimension-dependent factor of d.

This example highlights the capability of our method to efficiently handle composite structures that are naturally smooth in non-Euclidean geometries, avoiding unfavorable dimension scaling.