Mathematical Programming
https://doi.org/10.1007/s10107-025-02313-1

FULL LENGTH PAPER

Series A

Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions

Teodor Rotaru[1,2] · **François Glineur**[2] · **Panagiotis Patrinos**[1]

Received: 16 August 2024 / Accepted: 29 November 2025

Abstract

We consider gradient descent with constant stepsizes and derive exact worst-case convergence rates on the minimum gradient norm of the iterates. Our analysis covers all possible stepsizes and arbitrary upper/lower bounds on the curvature of the objective function, thus including convex, strongly convex and weakly convex (hypoconvex) objective functions. Among the challenging parts of the analysis, we note the necessity to exploit dependencies between non-consecutive iterates. While this complicates the proofs to some extent, it enables us to achieve an exact full-range analysis of gradient descent for any constant stepsize (covering, in particular, normalized stepsizes greater than one), whereas the literature contained only conjectured rates of this type. In the nonconvex case, allowing arbitrary bounds on upper and lower curvatures extends existing partial results that are valid only for gradient Lipschitz functions (i.e., where lower and upper bounds on curvature are equal), leading to improved rates for weakly convex functions. From our exact worst-case performance bounds, we deduce the optimal constant stepsize for gradient descent. Leveraging our analysis, we also introduce a new variant of gradient descent based on a unique, fixed sequence of variable stepsizes, demonstrating its superiority in the worst-case over any constant stepsize schedule.

Keywords Gradient descent · Performance estimation problem · Worst-case complexity analysis · Convex minimization · Weakly convex optimization

✉ Teodor Rotaru
teodor.rotaru2605@gmail.com

François Glineur
francois.glineur@uclouvain.be

Panagiotis Patrinos
panos.patrinos@esat.kuleuven.be

[1] Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Kasteelpark Arenberg 10, Leuven 3001, Belgium

[2] Department of Mathematical Engineering (ICTEAM-INMA), UCLouvain, Av. Georges Lemaître 4, Louvain-la-Neuve 1348, Belgium

Published online: 26 May 2026

Mathematics Subject Classification 65K15 · 90C25 · 90C26 · 90C30 · 49M37

1 Introduction

Tight convergence rates for the canonical first-order optimization algorithm, i.e., the gradient descent method, applied to smooth functions are demonstrated. We analyse functions with arbitrary upper and lower curvatures, covering weakly convex, convex and strongly convex cases. For all constant stepsizes ensuring descent, we derive tight bounds and establish convergence rates parameterized by curvature bounds, stepsize, and iteration count. The key tool enabling to obtain a complete characterization is the analysis of iterations' interconnection not only for consecutive steps, but also for the ones lying at distance two, unlike classical convergence analysis, e.g., [6, 7, 27]. Our proof strategy differs from existing work on performance estimation problems (PEP) (e.g., [22]), which connects consecutive iterates to the final one. In contrast, our approach allows for an analysis decoupled from the last iterate, using standard descent lemmas to establish performance bounds. Results on full range of stepsizes were previously at most conjectured for the apparently simpler scenario of convex functions.

1.1 Notations and definitions

We consider the unconstrained optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x),$$

where f belongs to the class of functions $\mathcal{F}_{\mu,L}$, as defined below.

Definition 1.1 Let $L \in (0,\infty)$ and $\mu \in (-\infty, L)$. Then a function $f\colon \mathbb{R}^d \to \mathbb{R}$ belongs to the class of functions $\mathcal{F}_{\mu,L}$ if it has:

(i) upper curvature L, i.e., $\frac{L}{2}\|\cdot\|^2 - f$ is convex;
(ii) lower curvature μ, i.e., $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Intuitively, if $f \in \mathcal{C}^2$, then the eigenvalues of its Hessian matrix belong to some interval $[\mu, L]$. Depending on the sign of μ, f is categorized as: (i) weakly convex (or hypoconvex) for $\mu < 0$, (ii) convex for $\mu = 0$ or (iii) strongly convex for $\mu > 0$. The trivial case $\mu = L$ is not explicitly addressed. In our framework, we allow $\mu < -L$, implying that the Lipschitz constant of the gradients is $\max\{-\mu, L\}$. The class of smooth weakly convex functions includes the smooth non-necessarily convex ones under the condition $\mu = -L$. We sometimes represent the curvature ratio $\kappa := \frac{\mu}{L} \in (-\infty, 1)$, which, if $\mu \geq 0$, signifies the inverse condition number.

Gradient descent. Starting from some initial point $x_0 \in \mathbb{R}^d$, we consider N iterations of gradient descent with fixed stepsizes $\gamma_i \in (0, \frac{2}{L})$, defined as:

$$x_{i+1} = x_i - \gamma_i \nabla f(x_i), \quad \forall i = 0, \ldots, N-1. \tag{GD}$$

The constraint on the stepsize values ensures a decrease in the function value after a single iteration. We assume access to a first-order oracle denoted by the set of triplets $\mathcal{T} := \{(x_i, g_i, f_i)\}_{i\in\mathcal{I}} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$, where $\mathcal{I}$ is an index set, providing information about the iterates, their gradients and the corresponding function values. We assume the function f to be bounded from below and denote $f_* := \inf_x f(x) > -\infty$. In the nonconvex case, where first-order methods cannot guarantee convergence to a global minimizer, analysis focuses on the stationarity measure $\min_{0\leq i\leq N}\{\|\nabla f(x_i)\|^2\}$, typically bounded in terms of the initial optimality global gap $f(x_0) - f_*$, or, when f_* is unknown, by the observable decrease $f(x_0) - f(x_N)$.

We denote $[x]_+ := \max\{x, 0\}$ for any $x \in \mathbb{R}$.

Remark 1.2 A tight (exact) performance bound is established in two steps: (i) providing an *upper bound* of the convergence measure and (ii) constructing a problem instance (a *lower bound*) achieving it. For convenience, the performance measure is often called *(convergence) rate*. This term refers not only to the asymptotic expression, such as $\mathcal{O}(\cdot)$, but also to its associated, and often exact, constant.

1.2 Prior work

The conventional analysis of gradient descent relies on a constant stepsize schedule and provides non-tight rates (for example, see the textbooks [5–8, 27]). Drori and Teboulle's seminal paper on performance estimation problem (PEP) [16] propelled the quest for precise convergence rates in first-order optimization methods, due to its systematic approach to analyze worst-case behaviors. The PEP frames the task of identifying an algorithm's most unfavorable behavior within a specific problem class as an optimization problem.

Using relaxation techniques, they formulate the problem as a convex semidefinite program. Taylor et al. [30, 32] advanced the framework by specifying conditions on when the relaxation is tight. Notably, they introduced interpolation inequalities as a necessary and sufficient tool for deriving exact performance bounds. The conditions they introduced for $\mathcal{F}_{\mu,L}$ functions are central to our derivations.

Convergence rates of first-order methods applied to smooth convex functions have been extensively examined using PEP. For gradient descent, in [16] tight rates are proved for smooth convex functions with stepsizes $\gamma L \in (0, 1]$ and conjectured for $\gamma L \in (1, 2)$; exact rates for smooth strongly convex functions are conjectured for $\gamma L \in (0, 2)$ [32] and proved when employing line-search [13]. Optimized first-order methods can be derived within the PEP framework, e.g., in [22] for an optimal algorithm to decrease the gradient norm of smooth convex functions. A recent breakthrough by Teboulle [34] establishes tight convergence rates for smooth convex functions using stepsizes $\gamma L \in (1, \frac{3}{2}]$.

For smooth nonconvex functions, commonly encountered in fields like machine learning, there is a notable gap in comprehending convergence rates. Carmon et al. establish lower bounds for finding stationary points in [9, Theorem 2], demonstrating that gradient descent is worst-case optimal for functions with only a Lipschitz gradient. In a related study, Cartis et al. present tight lower bounds for steepest descent [10]. Within the PEP framework, Taylor examines the convergence rates of gradient descent

applied to smooth non-necessarily convex functions with a constant stepsize $\frac{1}{L}$ [29, page 190], Drori and Shamir extend this result for stepsizes below $\frac{1}{L}$ [15, Corollary 1], while Abbaszadehpeivasti et al. further improve upon these findings [1], achieving state-of-the-art for stepsizes up to $\frac{\sqrt{3}}{L}$. Typically, these aforementioned results solely rely exploiting smoothness, neglecting the weak convexity parameter μ. Recent interest in weakly convex optimization has emerged, e.g., [4, 12].

Given our focus on constant stepsize schedule, our analysis is confined to the interval $\gamma L \in (0, 2)$. However, recent advancements address longer stepsizes and achieve improved asymptotic rates. Interested readers are directed to the works of Altschuler and Parrilo [2, 3], Das Gupta et al. [11] and Grimmer et al. [19–21]. Notably, similar to our approach, proving their rates requires connecting more than just consecutive iterations, deviating from classical analyses. Comparable demonstrations are provided in [22, 38].

Nesterov highlighted the significance of minimizing the gradient norm [26], asserting that $\|\nabla f(x)\|$ might serve better than other performance metrics for minimization purposes. Evaluating the gradient norm for initially bounded functions is a common approach in analyzing gradient-based methods for smooth nonconvex optimization, seen in works such as [15] for stochastic gradient descent and in [22] for the OGM-G algorithm optimizing the gradient norm for smooth convex functions. This convergence measure is suitable for applications where both the initial optimality gap and the gradient norm across iterations can be directly measured, unlike metrics involving the optimizer. We maintain consistency by utilizing this performance metric for both convex and strongly convex functions.

1.3 Paper organization and main contributions

Our contributions can be summarized as follows:

1. We provide tight performance bounds for gradient descent with constant stepsizes $\gamma L \in (0, 2)$, proving their tightness through worst-case examples. Our analysis applies to weakly convex (Theorem 2.4), strongly convex (Theorem 2.3), and convex functions (Theorem 2.2), with a comprehensive summary in Table 1.
2. A performance bound for non-necessarily convex functions using variable stepsizes $\gamma_i L \in (0, \overline{\gamma L}_1(\kappa)]$ (Theorem 2.5), which we prove is tight for stepsizes below 1 and conjecture for the remaining range. This result generalizes the constant stepsize case for weakly convex functions (Theorem 2.4).
3. Leveraging our tight performance bounds, we propose optimized stepsize rules for gradient descent, namely: (i) best constant stepsize for convex, strongly convex and weakly convex functions (Propositions 2.13 to 2.15) for a given number of iterations; (ii) a dynamic sequence of stepsizes, independent of the iteration count N, achieving a superior guarantee compared to any constant stepsize policy (Theorem 2.18).

The aforementioned works typically address only the convex case ($\mu = 0$), the strongly convex case ($\mu \in (0, L)$), the Lipschitz smooth case ($\mu = -L$), or the setting without lower curvature information ($\mu = -\infty$) (see, e.g., [27, Section 1.2.3]). Our framework

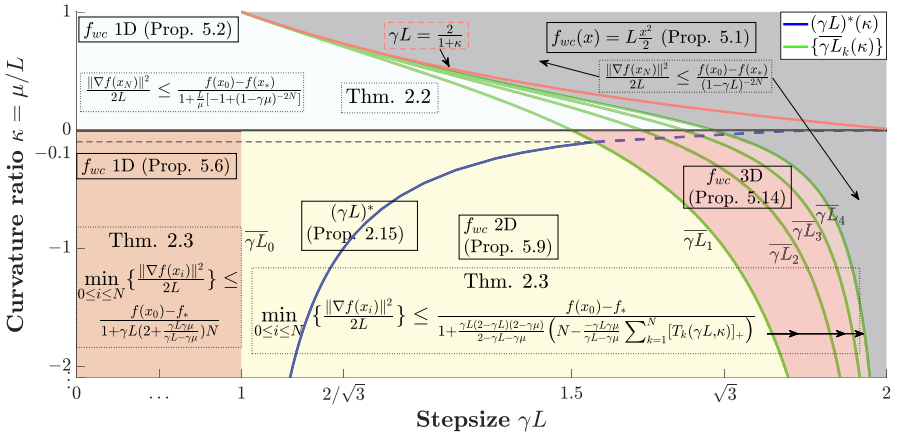


Fig. 1 Regimes outlined in Theorem 2.2 (smooth convex), Theorem 2.3 (smooth strongly convex) and Theorem 2.4 (smooth weakly convex) with respect to the constant stepsize γL and curvature ratio κ, for $N = 4$ iterations. State-of-the-art *proven* tight convergence rates are limited to the cases $\kappa = -\infty$ [27], $\kappa = -1$ [1] and $\kappa = 0$ [16, 34] and stepsizes lower than $\overline{\gamma L}_1(\kappa)$. Above this threshold, a series of N regimes parameterized by $k = 1, \ldots, N$ exist and they are sublinear for non-strongly convex functions. The optimal constant stepsize $(\gamma L)^*(\kappa)$ for weakly convex functions is independent of N for $\kappa \lesssim -0.1$, otherwise it is only asymptotically valid (as marked with *blue* dashed line). We use inequalities connecting three consecutive iterations to prove the regimes corresponding to stepsizes $\gamma L \in (\overline{\gamma L}_1, \frac{2}{1+[\kappa]_+})$. The worst-case function (f_{wc}) dimensionality and the tightness' proof references are illustrated.

unifies these regimes and yields a continuous spectrum of performance guarantees interpolating between them. We recall that the performance metric for (strongly) convex functions usually differs, e.g., in [16, Conjecture 1], [32, Conjecture 3] and [33, Theorem 2.1].

The gist in proving tight rates on the full interval of stepsizes is its partition into a collection of intervals delimited by stepsize thresholds dependent on the curvature ratio κ, each of them requiring distinct analysis. Figure 1 illustrates this partitioning for curvature ratio and (normalized) stepsizes, along with their corresponding rates for $N = 4$ iterations.

The remainder of the paper is organized as follows. Section 2 contains the formal statements of our main theoretical findings. Section 3 provides preliminary results required for the proofs. The complete proofs for the upper convergence bounds are detailed in Section 4, while Section 5 presents the worst-case examples that establish their tightness. To facilitate reproducibility, we provide a supplementary GitHub repository[1] containing MATLAB scripts for numerical verification and symbolic validation of the primary algebraic manipulations.

[1] GitHub repository: https://github.com/teo2605/GD_tight_rates

2 Performance bounds

In Sections 2.1 to 2.3 we provide the exact performance bounds for constant stepsize schedules. Table 1 summarizes the different results involved in their proofs. In Section 2.4 we give the tight bounds for smooth non-necessarily convex functions when using variable stepsizes, upper bounded by some threshold. Section 2.5 describes the challenging stepsize thresholds defining the regimes for constant stepsizes. Section 2.6 provides optimal stepsize rules with respect to the theoretical worst-case scenarios: best constant stepsize policies in Section 2.6.1 and a dynamic stepsize schedule with better worst-case guarantees in Section 2.6.2.

The convergence rates are influenced by: (i) curvature ratio $\kappa = \frac{\mu}{L}$, (ii) the (normalized) fixed stepsizes $\gamma_i L \in (0, 2)$, and (iii) the number of iterations N. Given a *fixed budget* (number of iterations N), the rates show stepsize dependent accuracies in finding a stationary point, of the type

$$\frac{1}{2L} \min_{0 \leq i \leq N} \left\{ \|\nabla f(x_i)\|^2 \right\} \leq \frac{f(x_0) - f_*}{1 + \tau_N(\kappa, (\gamma_i L)_{i=0}^{N-1})}, \tag{$\mathcal{P}_*$}$$

where $\tau_N(\cdot)$ is some function we determine through our analysis.

Remark 2.1 To be consistent with standard performance bounds in the literature, our final theorems use the optimality gap $f(x_0) - f_*$, as in ($\mathcal{P}_*$). We derive these results by providing a bound on the gap $f(x_0) - f(x_N)$, which is a measure directly available from the first-order oracle:

$$\frac{1}{2L} \min_{0 \leq i \leq N} \left\{ \|\nabla f(x_i)\|^2 \right\} \leq \frac{f(x_0) - f(x_N)}{\tau_N(\kappa, (\gamma_i L)_{i=0}^{N-1})}. \tag{$\mathcal{P}_N$}$$

The two bounds differ only by a "+1" term in the denominator. In nonconvex settings, e.g., multimodal functions, possessing multiple local optima, ($\mathcal{P}_N$) can be more effective for characterizing local convergence as it avoids the distortion of the numerator from a distant global optimum f_* with a much lower value of f_*.

2.1 Performance bounds for convex functions

Theorem 2.2 *Let $f \in \mathcal{F}_{0,L}$ and consider N iterations of* (GD) *with constant stepsizes $\gamma L \in (0, 2)$ starting from x_0. Then the following bound holds:*

$$\frac{1}{2L} \|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f_*}{1 + \gamma L \min\left\{2N\,,\ \frac{-1 + (1 - \gamma L)^{-2N}}{\gamma L}\right\}}. \tag{2.1}$$

In Proposition 5.4 we provide a worst-case function attaining the performance bound (2.1). To our knowledge, this result establishes the first *proved* tight rate in the convex case for the full range of constant stepsizes. The denominator in (2.1) is similar to the numerically conjectured one from [16, Conjecture 3.1], even though for a different

Table 1 Overview of performance bounds proofs and their tightness for all regimes with constant stepsizes (the thresholds $\overline{\gamma L}_k$, formally defined in Section 2.5, increase with k). The demonstrations use interpolation inequalities connecting consecutive iterations (distance-1) or, additionally, the ones situated at distance-2. Tightness is showed by constructing a worst-case function or by proving its existence via interpolating triplets.

Class	Stepsizes	Key result	Inequalities	Tightness proof	Worst-case example
convex $\mu = 0$ Theorem 2.2	$(0, \frac{3}{2}]$	Lemma 4.2 & Lemma 4.3	distance-1	Proposition 5.4	1D function
	$(\frac{3}{2}, \overline{\gamma L}_N)$	Corollary 4.14	distance-1&2		
	$[\overline{\gamma L}_N, 2)$	Lemma 4.12	distance-1&2	Proposition 5.2	
strongly convex $\mu > 0$ Theorem 2.3	$(0, \overline{\gamma L}_1]$	Lemma 4.8	distance-1	Proposition 5.3	1D function
	$[\overline{\gamma L}_1, \overline{\gamma L}_{N-1})$	Lemma 4.8 & Lemma 4.13	distance-1&2		
	$[\overline{\gamma L}_{N-1}, \overline{\gamma L}_N)$	Lemma 4.13	distance-1&2		
	$[\overline{\gamma L}_N, \frac{2}{1+\kappa})$	Lemma 4.12	distance-1&2	Proposition 5.2	
	$[\frac{2}{1+\kappa}, 2)$	Lemma 4.7	distance-1		
weakly convex $\mu < 0$ Theorem 2.4	$(0, 1]$	Lemma 4.2	distance-1	Proposition 5.7	1D function
	$[1, \overline{\gamma L}_1]$	Lemma 4.3	distance-1	Proposition 5.10	2D triplets
	$[\overline{\gamma L}_1, \overline{\gamma L}_{N-1})$	Lemma 4.3 & Lemma 4.11	distance-1&2	Proposition 5.14	3D triplets
	$[\overline{\gamma L}_{N-1}, \overline{\gamma L}_N)$	Lemma 4.11	distance-1&2	Proposition 5.12	2D function
	$[\overline{\gamma L}_N, 2)$	Lemma 4.12	distance-1&2	Proposition 5.2	1D function

performance metric, with $x_* \in \arg\min_x f(x)$:

$$f(x_N) - f_* \leq \frac{L}{2} \frac{\|x_0 - x_*\|^2}{1 + \gamma L \min\left\{2N\,,\ \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}\right\}}. \tag{2.2}$$

2.2 Performance bounds for strongly convex functions

Theorem 2.3 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (0, L)$, and consider N iterations of* (GD) *with stepsizes $\gamma L \in (0, 2)$ starting from x_0. Then the following bound holds:*

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f_*}{1 + \gamma L \min\left\{\frac{-1+(1-\gamma\mu)^{-2N}}{\gamma\mu}\,,\ \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}\right\}}. \tag{2.3}$$

We show the tightness of the bound (2.3) with a constructive example in Proposition 5.3. Tight performance bounds for stepsizes below $\frac{2}{L+\mu}$ are given in [32, Table 2][2]. For the metric analysed in this paper, however, only the rate corresponding to the stepsize $\gamma = \frac{1}{L}$ is conjectured there. In [30], the tight contraction factor $\max\{\gamma L - 1, 1 - \gamma\mu\}$ is derived and is optimized by choosing $\gamma = \frac{2}{L+\mu}$. Notably, this stepsize choice is suboptimal for the metric considered in our work (see Proposition 2.14) and is also outperformed by the dynamic stepsize procedure proposed in Section 2.6.2. Subsequent to the publication of our work on *arXiv*, [23] provided an alternative proof of Theorem 2.3. In the same paper, exploiting the *H-duality* phenomenon [24], the author provides the first proof for the numerically conjectured rate from [32, Conjecture 2], which in our notation writes as:

$$f(x_N) - f_* \leq \frac{L}{2} \frac{\|x_0 - x_*\|^2}{1 + \gamma L \min\left\{\frac{-1+(1-\gamma\mu)^{-2N}}{\gamma\mu}\,,\ \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}\right\}}, \tag{2.4}$$

where $x_* = \arg\min_x f(x)$. Letting $\mu \searrow 0$, the rate for convex functions from Theorem 2.2 is recovered.

2.3 Performance bounds for weakly convex functions

Theorem 2.4 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu < 0$, and consider N iterations of* (GD) *with constant stepsizes $\gamma L \in (0, 2)$ starting from x_0. Then the following bound holds:*

$$\frac{1}{2L} \min_{0\leq i\leq N} \left\{\|\nabla f(x_i)\|^2\right\} \leq \frac{f(x_0) - f_*}{1 + \gamma L \min\left\{P_N(\gamma L, \gamma\mu)\,,\ \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}\right\}}, \tag{2.5}$$

[2] The work in [32] provides tight performance bounds for the proximal gradient method, which generalizes gradient descent.

with $P_N(l,u)$ defined as:

$$P_N(l,u) := p(l,u)\left[N - \frac{-lu}{l-u}\sum_{k=1}^{N}\left[\frac{-1+(1-u)^{-2k}}{u} - \frac{-1+(1-l)^{-2k}}{l}\right]_+\right] \tag{2.6}$$

where

$$p(l,u) := 2 - \frac{-lu}{1-u-|1-l|} = \begin{cases} 2 - \frac{-lu}{l-u}, & l \in (0,1]; \\ \frac{(2-l)(2-u)}{2-l-u}, & l \in [1,2). \end{cases} \tag{2.7}$$

We establish the tightness of these rates over the entire stepsize interval. Explicit worst-case functions are provided for $\gamma L \in (0,1]$ (one-dimensional, Proposition 5.7), $\gamma L \in [\overline{\gamma L}_{N-1}(\kappa), \overline{\gamma L}_N(\kappa))$ (two-dimensional, Proposition 5.12) and $\gamma L \in [\overline{\gamma L}_N(\kappa), 2)$ (one-dimensional, Proposition 5.2). For the remaining intervals, the worst-case behaviour is captured by interpolating triplets: $\gamma L \in (1, \overline{\gamma L}_1(\kappa)]$ (two-dimensional, Proposition 5.10) and $\gamma L \in [\overline{\gamma L}_1(\kappa), \overline{\gamma L}_N(\kappa))$ (three-dimensional, Proposition 5.14).

The rate's denominator combines a set of N sublinear regimes with a linear regime also observed for convex and strongly convex functions (Theorem 2.2 and Theorem 2.3), holding for stepsizes close to 2. For $\gamma L \in (0,1]$, each term in the sum appearing in (2.6) is equal to zero. With $\gamma L \in (1,2)$, only the terms up to some index k become positive, while the remaining $(N-k)$ terms are negative. The expression is continuous with respect to the stepsize. A new term in the sum becomes nonnegative when evaluated at specific stepsize thresholds $\overline{\gamma L}_k$ rigorously defined in Section 2.5 and satisfying the condition:

$$\frac{-1+(1-\kappa\overline{\gamma L}_k)^{-2k}}{\kappa\overline{\gamma L}_k} - \frac{-1+(1-\overline{\gamma L}_k)^{-2k}}{\overline{\gamma L}_k} = 0. \tag{2.8}$$

Using the identity

$$\sum_{j=1}^{k}\frac{-1+(1-u)^{-2j}}{u} - \frac{-1+(1-l)^{-2j}}{l} = \frac{-1+(1-u)^{-2k}}{1-(1-u)^2} - \frac{-1+(1-l)^{-2k}}{1-(1-l)^2} + \left(\frac{1}{l}-\frac{1}{u}\right)k,$$

the expression of $P_N(\gamma L, \gamma\mu)$ from (2.6) can be equivalently written as

$$P_N(\gamma L,\gamma\mu) = p(\gamma L,\gamma\mu)\min_{0\le k\le N}\left\{\frac{\frac{-1+(1-\gamma L)^{-2k}}{\gamma L[1-(1-\gamma L)^2]} - \frac{-1+(1-\gamma\mu)^{-2k}}{\gamma\mu[1-(1-\gamma\mu)^2]}}{\frac{1}{\gamma L}-\frac{1}{\gamma\mu}} + N - k\right\}.$$

For stepsize thresholds satisfying (2.8), it reduces to summing an exponential term with a linear one:

$$P_N(\overline{\gamma L}_k, \kappa\overline{\gamma L}_k) = \frac{(2-\overline{\gamma L}_k)(2-\kappa\overline{\gamma L}_k)}{2-\overline{\gamma L}_k-\kappa\overline{\gamma L}_k}(N-k) + \frac{-1+(1-\overline{\gamma L}_k)^{-2k}}{\overline{\gamma L}_k}.$$

When reaching the maximum threshold $\overline{\gamma L}_N$ corresponding to $k = N$, $P_N(\gamma L, \gamma\mu)$ becomes $\frac{-1+(1-\overline{\gamma L}_N)^{-2N}}{\overline{\gamma L}_N}$ and the rate shifts to the regime linear in $(1-\overline{\gamma L}_N)$. Summarizing, the expression of $P_N(\gamma L, \gamma\mu)$ can be expanded as follows:

$$\begin{cases} \left(2+\frac{\gamma\mu\gamma L}{\gamma L-\gamma\mu}\right)N, & \gamma L \in (0,1]; \\ \frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}(N-k) + \frac{\frac{-1+(1-\gamma L)^{-2k}}{\gamma L[1-(1-\gamma L)^2]} - \frac{-1+(1-\gamma\mu)^{-2k}}{\gamma\mu[1-(1-\gamma\mu)^2]}}{\frac{1}{1-(1-\gamma L)^2}-\frac{1}{1-(1-\gamma\mu)^2}}, & \gamma L \in \left[\overline{\gamma L}_k, \overline{\gamma L}_{k+1}\right) \\ & k=0,\ldots,N-1; \\ \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}, & \gamma L = \overline{\gamma L}_N. \end{cases}$$

By fixing stepsize γL and varying N, another perspective on convergence rates emerges: determining the maximum number of iterations to reach a desired accuracy with a *given stepsize*. Let $\bar{N}$ be the largest index k corresponding to a strictly positive term in expression of $P_N(\gamma L, \gamma\mu)$ from (2.6). Then for $N \leq \bar{N}$ the rate is linear in $(1-\gamma L)$, whereas for $N \geq \bar{N}+1$ it shifts to a sublinear regime.

2.4 Performance bounds for non-strongly convex functions and variable stepsizes

Theorem 2.5 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, 0]$, and consider N iterations of (GD) starting from x_0 with stepsizes $\gamma_i L \in (0, \overline{\gamma L}_1(\kappa)]$, $i = 0,\ldots,N-1$, where $\overline{\gamma L}_1(\kappa) := \frac{3}{1+\kappa+\sqrt{1-\kappa+\kappa^2}} \in [\frac{3}{2}, 2)$. Then the following bound holds, where $p(\gamma L, \gamma\mu)$ is defined in (2.7):*

$$\frac{1}{2L}\min_{0\leq i\leq N}\left\{\|\nabla f(x_i)\|^2\right\} \leq \frac{f(x_0)-f_*}{1+\sum_{i=0}^{N-1}\gamma_i L\, p(\gamma_i L, \gamma_i\mu)}. \tag{2.9}$$

The bound in (2.9) is proved to be tight for $\gamma_i L \in (0,1]$ in Proposition 5.7. For the range $\gamma_i L \in (1, \overline{\gamma L}_1(\kappa)]$, we conjecture its tightness based on a two-dimensional worst-case example (Conjecture 1).

Remark 2.6 Restricting it to constant stepsizes $\gamma L \leq \overline{\gamma L}_1(\kappa)$, Theorem 2.5 recovers the result from Theorem 2.4 corresponding to $\frac{-1+(1-\gamma\mu)^{-2k}}{\gamma\mu} - \frac{-1+(1-\gamma L)^{-2k}}{\gamma L} \leq 0$ for all $k \geq 1$ in (2.6), namely $P_N(\gamma L, \gamma\mu) = p(\gamma L, \gamma\mu)\, N$. The extension to variable stepsizes relies on the proof which, within the stepsize range from Theorem 2.5, only requires inequalities relating consecutive iterations.

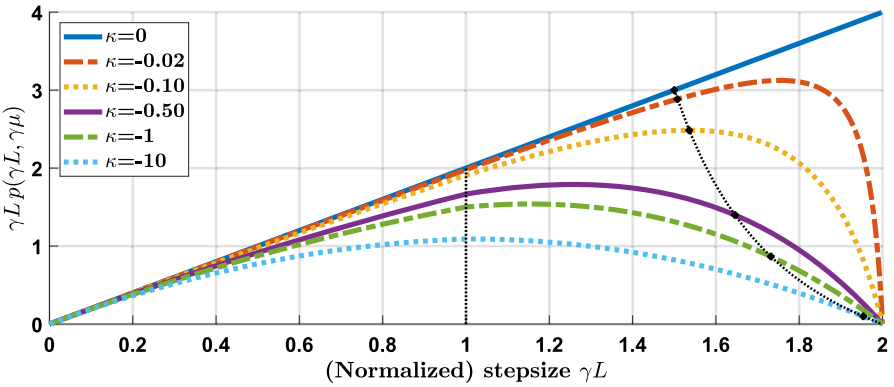


Fig. 2 The dominant term $\gamma L\, p(\gamma L, \gamma\mu)$ (see (2.7)) in the bounds for weakly convex functions for several curvature ratios. $\gamma L = 1$ delimits the two main sublinear regimes, while the one-step analysis is tight up to the threshold $\overline{\gamma L}_1(\kappa)$, marked in black dots, above which intervene transient contributions (see (2.6))

Figure 2 illustrates the behavior of the leading term $\gamma L\, p(\gamma L, \gamma\mu)$, emphasizing the stepsize threshold $\overline{\gamma L}_1(\kappa)$ and additionally showing that optimal rates, i.e., with larger denominators, are achieved in the regime of stepsizes greater than 1.

Remark 2.7 Theorem 2.5 extends to arbitrary $\mu < 0$ the state-of-the-art rate for smooth non-necessarily convex functions ($\mu = -L$) from [1, Theorem 2]:

$$\frac{1}{2L}\min_{0\le i\le N}\left\{\|\nabla f(x_i)\|^2\right\} \le \frac{f(x_0)-f_*}{1+\sum_{i=0}^{N-1}\gamma_i L\left(2-\frac{\gamma_i L}{2}\max\{1,\gamma_i L\}\right)},\ \forall \gamma_i L\in(0,\sqrt{3}]. \tag{2.10}$$

The convergence rate (2.11), stated, for instance, in [27, Section 1.2.3], is derived for smooth functions using only the upper curvature L:

$$\frac{1}{2L}\min_{0\le i\le N}\left\{\|\nabla f(x_i)\|^2\right\} \le \frac{f(x_0)-f_*}{1+\sum_{i=0}^{N-1}\gamma_i L(2-\gamma_i L)},\qquad \gamma_i L\in(0,2). \tag{2.11}$$

Theorem 2.4 recovers this result in the absence of lower curvature information, i.e., for $\mu \searrow -\infty$ and effectively becomes Theorem 2.5 covering the full stepsize domain $(0, 2)$ (as $\overline{\gamma L}_1(-\infty) = 2$). The proof of this celebrated result lacks the utilization of quadratic *lower* bounds and thus remains generally non-tight. The convergence rate from Theorem 2.4 exhibits a smooth interpolation between the result in (2.11) (when $\mu \searrow -\infty$), the rate for smooth non-necessarily convex functions ($\mu = -L$) from (2.10) and the rate on smooth convex functions ($\mu = 0$).

Corollary 2.8 *Let $f \in \mathcal{F}_{0,L}$ and consider N iterations of (GD) starting from x_0 with fixed stepsizes $\gamma_i L \in (0, \frac{3}{2}]$, $i = 0, \dots, N-1$. Then the following bound holds:*

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \le \frac{f(x_0)-f_*}{1+2\sum_{i=0}^{N-1}\gamma_i L}.$$

Corollary 2.8 is obtained by taking $\mu = 0$ in Theorem 2.5. Although not explicitly presented as such, Teboulle and Vaisbourd propose in [34] two sufficient decrease results (Lemma 1 and Lemma 5) yielding the same rate from Corollary 2.8 through telescoping summation (see also Remark 4.4).

2.5 Stepsize thresholds

Stepsize thresholds are given by the roots of (2.8). For N iterations, there are exactly N such thresholds, delimiting the regimes for $\gamma L > 1$. These thresholds play a central role in the proofs for (strongly) convex and weakly convex functions in Theorems 2.2 to 2.4, as they partition the stepsize domain into $N + 1$ distinct regimes, each corresponding to a different worst-case instance. We begin by introducing the following auxiliary expressions.

Definition 2.9 Let k be a positive integer and define function $E_k : \mathbb{R}\backslash\{0\} \to \mathbb{R}$:

$$E_k(x) := \sum_{j=1}^{2k} x^{-j} = \begin{cases} \frac{-1+x^{-2k}}{1-x}, & x \neq 1; \\ 2k, & x = 1. \end{cases} \tag{E_k}$$

Let $V := \{(\ell, \kappa) : \kappa \in (-\infty, 1),\ \ell \in (1, \frac{2}{1+[\kappa]_+})\}$ and define $T_k : V \to \mathbb{R}$ as

$$T_k(\ell, \kappa) := E_k(1 - \kappa\ell) - E_k(1 - \ell). \tag{T_k}$$

$T_k(\gamma L, \kappa)$ denotes exactly the individual terms summed up in (2.6) and canceled by the k-th stepsize threshold in (2.8). Moreover, $T_N(\gamma L, \kappa)$ is the difference of the arguments in the denominators of the rates for convex functions (2.1) and strongly convex functions (2.3), respectively.

Proposition 2.10 *The function $T_k(\cdot, \kappa)$ is strictly increasing in the first argument.*

Proof See Appendix A.1. □

The sign of $T_k(\gamma L, \kappa)$ changes in the stepsize interval $\gamma L \in \left(1, \frac{2}{1+[\kappa]_+}\right)$, on which it therefore possesses a unique root.

Definition 2.11 (Stepsize thresholds) Let $\overline{\gamma L}_\infty(\kappa) := \frac{2}{1+[\kappa]_+}$. The unique roots of $T_k(\gamma L, \kappa)$ in the interval $(1, \overline{\gamma L}_\infty(\kappa))$ are referred to as stepsize thresholds and denoted by $\overline{\gamma L}_k(\kappa)$, where $\overline{\gamma L}_0(\kappa) := 1$ and:

$$\left\{\overline{\gamma L}_k(\kappa)\right\} := \left\{\gamma L \in (1, \overline{\gamma L}_\infty(\kappa)) \mid T_k(\gamma L, \kappa) = 0\right\} \quad \forall k = 1, 2, \ldots . \tag{$\overline{\gamma L}_k$}$$

The definition of stepsize thresholds ($\overline{\gamma L}_k$) resembles exactly the condition (2.8) marking the transitions between the different regimes indexed by k in Theorem 2.4.

Proposition 2.12 *The stepsize thresholds $\overline{\gamma L}_k(\kappa)$ satisfy the following properties:*
1. $\overline{\gamma L}_k(\kappa) < \overline{\gamma L}_{k+1}(\kappa)$, for all integers $k \geq 0$;

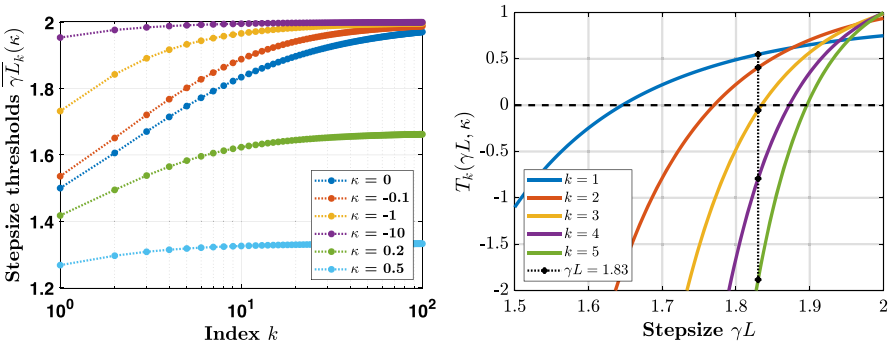


Fig. 3 Stepsize thresholds and the sign transition of T_k as a function of the normalized stepsize γL, for fixed curvature ratio κ. (**a**) Stepsize thresholds $\overline{\gamma L}_k(\kappa)$ converge to $\frac{2}{1+\kappa}$ for strongly convex functions and to 2 for non-strongly convex functions. (**b**) Dependence of $T_k(\gamma L, \kappa)$ on $\gamma L (\kappa = -0.5)$. For $\gamma L = 1.83$, $T_k \geq 0$ for $k \in \{1, 2\}$; $\overline{\gamma L}_k(\kappa)$ mark the x-intercept

2. $\lim_{k\to\infty} \overline{\gamma L}_k(\kappa) = \overline{\gamma L}_\infty(\kappa)$*;*
3. For any $\gamma L \in \left[\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa)\right)$*:* $T_i(\gamma L, \kappa) \geq 0\ \forall i \leq k$*, and* $T_i(\gamma L, \kappa) < 0$ $\forall i \geq k+1$.

Proof See Appendix A.2. □

Figure 1 illustrates the dependence on κ of the first $N = 4$ stepsize thresholds. Complementary, Figure 3a details the rapid increase of the thresholds with the number of iterations, reaching: (i) $\frac{2}{1+\kappa}$ for strongly convex functions and (ii) 2 in the non-strongly convex case. The distance between two consecutive thresholds decreases when increasing their index. Figure 3b depicts $T_k(\gamma L, \kappa)$ defined in (T_k) for $\kappa = -0.5$ and several indices k, offering insight on how to determine the stepsize thresholds. For example, with $\gamma L = 1.83$, $T_k(\gamma L, \kappa)$ is positive for $k \leq 2$ and negative for $k \geq 3$; thus, the interval corresponding to these conditions is $[\overline{\gamma L}_2, \overline{\gamma L}_3)$.

2.6 Improved stepsize schedules

Building on our worst-case performance analysis, we provide several stepsize policies that minimize these upper bounds. Recommendations are provided for both constant (Section 2.6.1) and variable stepsizes (Section 2.6.2).

2.6.1 Optimal constant stepsizes

Proposition 2.13 *(Constant stepsize for convex functions) Let* $f \in \mathcal{F}_{0,L}$ *and consider* N *iterations of* (GD) *starting from* x_0 *with constant stepsizes. The worst-case rate for convex functions* (2.1) *is minimized by the choice* $\gamma L = \overline{\gamma L}_N(0)$*, implying the best worst-case bound*

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f_*}{1 + 2N\ \overline{\gamma L}_N(0)}. \tag{2.12}$$

Proposition 2.14 *(Constant stepsize for strongly convex functions) Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (0, L)$, and N iterations of (GD) starting from x_0 with constant stepsizes. The worst-case rates for strongly convex functions (2.3) is minimized by the choice $\gamma L = \overline{\gamma L}_N(\kappa)$, implying the best worst-case bound*

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f_*}{(1 - \overline{\gamma L}_N(\kappa))^{-2N}}. \tag{2.13}$$

When evaluated at the stepsizes threshold $\overline{\gamma L}_N(\kappa)$, Definition 2.11 implies that the two arguments in the minimization defining the performance bounds of Theorem 2.2 and Theorem 2.3 coincide.

The same optimal constant stepsize $\overline{\gamma L}_N(\kappa)$ is minimizing the worst-case rate for the more standard performance criterion $\frac{f(x_N)-f_*}{\|x_0-x_*\|^2}$, since the denominators are the same (see Theorem 2.2 for convex functions and Theorem 2.3 for strongly convex functions). We refer the reader to the discussion following [16, Conjecture 3.1] for convex functions and [32, Sections 4.1.1-2] for strongly convex functions. These rates were recently proved in [23].

Proposition 2.15 *(Constant stepsize for weakly convex functions) Let $\kappa < 0$, $\bar{\kappa} := \frac{-9-5\sqrt{5}+\sqrt{190+90\sqrt{5}}}{4} \approx -0.1001$, and define $(\gamma L)^*(\kappa)$ as the unique solution belonging to the interval $[1, 2)$ of the equation:*

$$-\kappa(1+\kappa)(\gamma L)^3 + \left[3\kappa + (1+\kappa)^2\right](\gamma L)^2 - 4(1+\kappa)(\gamma L) + 4 = 0. \tag{2.14}$$

With respect to minimizing the tight performance bound for weakly convex functions from Theorem 2.4, $(\gamma L)^*(\kappa)$ is:

1. *the optimal constant stepsize for any N, when $\kappa \leq \bar{\kappa}$;*
2. *the asymptotically optimal constant stepsize as $N \to \infty$, i.e., when the algorithm runs for a sufficiently large number of iterations, if $\kappa \in (\bar{\kappa}, 0)$.*

Proof See Appendix A.2. □

Unlike the optimal stepsize $\frac{2}{\sqrt{3}L}$ provided in [1, Theorem 3] for the usual smooth non-necessarily convex case $\mu = -L$, the recommendation in Proposition 2.15 leverages lower curvature information and maximizes the leading term $\gamma L\, p(\gamma L, \gamma\mu)$ in the rate's denominator from Theorem 2.4. As κ approaches $-\infty$, implying the absence of lower curvature information, the quantity $\gamma L(2 - \gamma L)$ is maximized, resulting in the celebrated optimal stepsize $\frac{1}{L}$. Similarly, substituting $\kappa = -1$ in equation (2.14) yields the optimal stepsize for smooth non-necessarily convex functions from [1].

Figure 1 depicts the continuous dependence of $(\gamma L)^*(\kappa)$ on the curvature ratio; above $\bar{\kappa} \approx -0.1$, the dashed line marks its asymptotic optimality. Figure 4a illustrates the maximum guaranteed improvement achievable by optimizing the leading term $p^{-1}(\gamma L, \gamma\mu)$ and compares: (i) the (asymptotically) optimal stepsize $(\gamma L)^*(\kappa)$ from Proposition 2.15, (ii) the optimal stepsize for smooth functions $(\gamma L)^*(-1) = \frac{2}{\sqrt{3}}$,

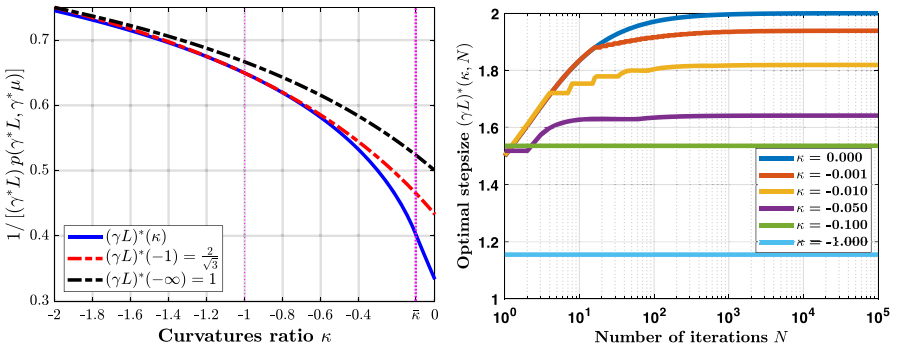


Fig. 4 Optimal constant stepsize selection for weakly convex functions. (**a**) Comparsion of (normalized) optimal stepsize recommendations: (i) $(\gamma L)^*(\kappa)$ from Proposition 2.15, (ii) $\gamma L = \frac{2}{\sqrt{3}}$ from [1], and (iii) $\gamma L = 1$ (classical). (**b**) Optimal stepsize versus N (log scale): asymptotic convergence to $(\gamma L)^*(\kappa)$; near convexity, a transient staircase behavior appears

and (iii) the celebrated optimal stepsize $(\gamma L)^*(-\infty) = 1$. As nonconvexity decreases ($\kappa \nearrow 0$), the optimal stepsize from Proposition 2.15 exhibits a greater improvement compared to stepsize $(\gamma L)^*(-1)$, which does not employ lower curvature information.

In contrast to the constant optimal stepsize recommendations for (strongly) convex functions, in the weakly convex case there exists an asymptotically optimal schedule independent of the number of iterations. A transient regime appears only for $-0.1 \lesssim \kappa < 0$, where the optimal constant stepsize also varies with N, i.e., $(\gamma L)^* = (\gamma L)^*(\kappa, N)$, increasing monotonically. Figure 4b shows this dependency for several curvature ratios, where the worst-case is numerically minimized for a given iteration budget. As $\kappa \nearrow 0$, the transient regime to reach the asymptotically optimal stepsize from Proposition 2.15 expands. During the initial iterations, a staircase-like behavior is observed and the optimal stepsize is confined within a specific subdomain determined by the stepsize thresholds, i.e., $(\gamma L)^*(\kappa, N) \in [1, \overline{\gamma L}_N(\kappa))$. In the convex case, the staircase behavior disappears, as the optimal constant stepsize $(\gamma L)^*(0, N)$ becomes $\overline{\gamma L}_N(0)$, lying at the intersection of two linear regimes (see Proposition 2.13).

2.6.2 Dynamic stepsizes

Stepsizes with $\gamma L > 1$ admit faster rates, motivating analysis in this regime. For (strongly) convex functions, variable stepsize policies can attain superior worst-case performance to constant ones by occasionally exceeding 2 [2, 3, 21, 39, 40], but they rely on intricate proofs, may depend on a fixed horizon, and their non-monotone nature makes them sensitive to curvature misestimation.

We introduce a monotonically increasing stepsize sequence inspired by Theorem 2.5 and the work of Teboulle-Vaisbourd [34], that avoids these issues. It applies to smooth strongly convex, convex, and weakly convex functions, is horizon-free, achieves better worst-case guarantees than the optimal constant stepsize, and admits significantly simpler convergence proofs. The construction is derived by telescoping

a two-point descent inequality (Lemma 4.3), yielding a recurrence that cancels a key term and produces the largest admissible stepsize at each iteration.

Definition 2.16 (Dynamic stepsizes) Let $\kappa \in (-\infty, 1)$. We consider the sequence $\{s_k(\kappa)\}_{k=-1}^{\infty}$, with $s_{-1} = 0$, recursively defined as:

$$s_{k+1}(\kappa) := \Big\{ s_+ \in \big(1, \tfrac{2}{1+[\kappa]_+}\big) : \frac{s_+[(2-s_+)(2-\kappa s_+)-1]}{2-s_+(1+\kappa)} + \frac{s_k(\kappa)}{2-s_k(\kappa)(1+\kappa)} = 0 \Big\}. \tag{2.15}$$

Proposition 2.17 *For any fixed $\kappa \in (-\infty, 1)$, the sequence $(s_k(\kappa))_{k=-1}^{\infty}$ is: (i) uniquely defined; (ii) monotonically increasing; and (iii) $\lim_{k\to\infty} s_k(\kappa) = \frac{2}{1+[\kappa]_+}$. Moreover, for any $k \geq 1$ it holds that (iv) $s_{k-1}(0) \geq 2 - \frac{1}{\frac{3}{2}k}$; (v) $s_{k-1}(\kappa) \geq \frac{2}{1+\kappa} - (\frac{1-\kappa}{1+\kappa})^{2k}$, if $\kappa \in (0, 1)$.*

Proof See Appendix A.3. □

Note that $s_0 = \overline{\gamma L}_1(\kappa)$ is the largest value for which Theorem 2.5 holds. Obtaining a closed-form expression for s_k is difficult, as each step requires solving a quadratic equation (in the convex case) or cubic one (when $\kappa \neq 0$).

Theorem 2.18 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, L)$, and $\kappa = \frac{\mu}{L}$. Consider N iterations of (GD) starting from x_0 with stepsizes $\gamma_i L$ defined as follows:*

$$\gamma_i L := \begin{cases} \min\{s_i(\kappa),\ (\gamma L)^*(\kappa)\}, & \kappa < 0 \\ s_i(\kappa), & \kappa \geq 0 \end{cases}, \quad \forall i = 0, \ldots, N-1,$$

where $(\gamma L)^*(\kappa)$ is defined in Proposition 2.15 and $s_i(\kappa)$ is given in Definition 2.16. Then the following bound holds:

$$\frac{1}{2L} \min_{0\leq i\leq N} \left\{ \|\nabla f(x_i)\|^2 \right\} \leq \frac{f(x_0) - f_*}{1 + \sum_{i=0}^{N-1} \frac{\gamma_i L(2-\gamma_i L)(2-\kappa\gamma_i L)}{2-\gamma_i L(1+\kappa)}}. \tag{2.16}$$

Proof See Appendix A.4. □

Based on numerical observations, we conjecture that the bound in (2.16) is tight. Theorem 2.18 extends Theorem 2.5 to larger stepsizes (i.e., $> \overline{\gamma L}_1(\kappa)$) and also applies to the strongly convex case ($\kappa > 0$). Theorem 2.18 is proved in Appendix A.4, together with the resulting corollaries for convex, strongly convex and weakly convex functions (Corollaries 2.19 to 2.21).

Corollary 2.19 *Let $f \in \mathcal{F}_{0,L}$ and consider N iterations of (GD) starting from x_0 with stepsizes $\gamma_i L = s_i(0) = 1 + \frac{2}{1+\sqrt{9-4s_{i-1}(0)}}$ for all $i = 0, \ldots, N-1$. Then the following bound holds:*

$$\frac{1}{2L} \|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f_*}{1 + \frac{s_{N-1}(0)}{2-s_{N-1}(0)}}. \tag{2.17}$$

Table 2 Comparison of complexity bounds (denominators) for convex functions between the policies: (i) celebrated stepsize $\gamma L = 1$; (ii) optimal constant stepsizes (Proposition 2.13) and (iii) dynamic schedule (from Corollary 2.19); higher is better. The last column displays the ratio of the these last two denominators.

N	Standard choice		Proposition 2.13		Corollary 2.19		
	$\gamma L = 1$	$1 + 2N$	$\overline{\gamma L}_N$	$1 + 2N\overline{\gamma L}_N$	s_{N-1}	$1 + \frac{s_{N-1}}{2-s_{N-1}}$	Ratio (%)
1	1	3	1.500	4.000	1.500	4.000	100.000
2	1	5	1.606	7.423	1.732	7.464	100.549
5	1	11	1.747	18.471	1.893	18.619	100.806
10	1	21	1.834	37.681	1.947	37.933	100.670
20	1	41	1.897	76.885	1.974	77.235	100.456
30	1	61	1.924	116.426	1.983	116.826	100.343
40	1	81	1.939	156.106	1.987	156.535	100.275
50	1	101	1.949	195.859	1.990	196.310	100.230
100	1	201	1.971	395.109	1.995	395.612	100.127

According to Proposition 2.17(iv), the bound can be relaxed by $\frac{f(x_0)-f_*}{3N}$, which recovers the standard rate of $\mathcal{O}(\frac{1}{N})$.

Corollary 2.20 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (0, L)$, and consider N iterations of (GD) starting from x_0 with stepsizes $\gamma_i L = s_i(\kappa)$ from Definition 2.16, with $i = 0, \ldots, N-1$. Then the following bound holds:*

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f_*}{1 + \frac{s_{N-1}(\kappa)}{2-s_{N-1}(\kappa)(1+\kappa)}}. \tag{2.18}$$

As a consequence of Proposition 2.17, the gradient norm converges to zero at a rate of $\mathcal{O}((\frac{1-\kappa}{1+\kappa})^{2N})$.

Following the approach in [34, Table 1], we numerically compare our dynamic stepsize policy with the optimal constant stepsize dependent on the number of iterations. This comparison is presented in Table 2 (convex case) and Table 3 (strongly convex case). Such a numerical comparison is necessary to evaluate the worst-case performance, as our dynamic policy lacks a closed-form expression.

Corollary 2.21 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, 0)$, and consider N iterations of (GD) starting from x_0 with stepsizes $\gamma_i L = \min\{s_i(\kappa)\,,\ (\gamma L)^*(\kappa)\}$, with $i = 0, \ldots, N-1$, $s_i(\kappa)$ defined in Definition 2.16 and $(\gamma L)^*(\kappa)$ given in Proposition 2.15. Then the following bound holds:*

$$\frac{1}{2L}\min_{0\leq i\leq N}\left\{\|\nabla f(x_i)\|^2\right\} \leq \frac{f(x_0) - f_*}{1 + p_* N + \max\limits_{0\leq k\leq N}\left\{\frac{s_{k-1}(\kappa)}{2-s_{k-1}(\kappa)\,(1+\kappa)} - k p_*\right\}}, \tag{2.19}$$

where $p_* := p((\gamma L)^*, \kappa(\gamma L)^*)$ is the maximized coefficient $p(\gamma L, \gamma\mu)$ from (2.7).

Table 3 Comparison of complexity bounds (denominators) for strongly convex functions between the policies: (i) celebrated stepsize $\gamma L = \frac{2}{1+\kappa}$; (ii) optimal constant stepsizes (Proposition 2.14) and (iii) dynamic schedule (Corollary 2.20); higher is better. The last column displays the ratio of these last two denominators.

κ	N	Standard choice		Proposition 2.14		Corollary 2.20		
		$\frac{2}{1+\kappa}$	$\left(\frac{1-\kappa}{1+\kappa}\right)^{-2N}$	$\overline{\gamma L}_N$	$(1-\overline{\gamma L}_N)^{-2N}$	s_{N-1}	$1+\frac{s_{N-1}}{2-s_{N-1}(1+\kappa)}$	Ratio (%)
10^{-3}	1	1.998	1.004	1.500	4.006	1.500	4.006	100.000
	2	1.998	1.008	1.605	7.447	1.731	7.488	100.551
	5	1.998	1.020	1.746	18.633	1.892	18.784	100.813
	10	1.998	1.041	1.833	38.381	1.946	38.643	100.682
	20	1.998	1.083	1.896	79.879	1.973	80.257	100.473
	30	1.998	1.127	1.923	123.416	1.982	123.865	100.363
	40	1.998	1.174	1.938	168.871	1.986	169.373	100.297
	50	1.998	1.221	1.948	216.257	1.989	216.805	100.253
	70	1.998	1.492	1.960	317.040	1.992	317.670	100.199
10^{-4}	1	1.9998	1.000	1.500	4.001	1.500	4.001	100.000
	2	1.9998	1.001	1.606	7.426	1.732	7.467	100.550
	5	1.9998	1.002	1.747	18.487	1.893	18.636	100.807
	10	1.9998	1.004	1.834	37.750	1.947	38.004	100.671
	20	1.9998	1.008	1.897	77.178	1.974	77.531	100.457
	30	1.9998	1.012	1.924	117.101	1.983	117.505	100.345
	40	1.9998	1.016	1.939	157.323	1.987	157.759	100.277
	50	1.9998	1.020	1.949	197.780	1.990	198.240	100.232
	70	1.9998	1.028	1.961	279.309	1.993	279.801	100.176

Table 4 Comparison of complexity bounds (denominators) for weakly convex functions (example $\kappa = -10^{-3}$) between the policies: (i) asymptotically constant optimal stepsize $(\gamma L)^*(\kappa)$ from Proposition 2.15, resulting in a denominator denoted by $P_N^A = P_N((\gamma L)^*(\kappa), \kappa(\gamma L)^*(\kappa))$; (ii) constant optimal stepsize depending on the number of iterations $(\gamma L)^*(N, \kappa)$ determined by numerically maximizing the denominator $P_N(\gamma L, \gamma\mu)$, denoted by $P_N^* = P_N\left((\gamma L)^*(N, \kappa), \kappa(\gamma L)^*(N, \kappa)\right)$; and (iii) dynamic sequence from Corollary 2.21, leading to the denominator D_N. The transition between the increasing sequence and the constant schedule is marked by the horizontal line. The last column displays the ratio $D_N(\kappa)/P_N^*$

N	Proposition 2.15		Proposition 2.15		Corollary 2.21		
	$(\gamma L)^*(\kappa)$	P_N^A	$(\gamma L)^*(N, \kappa)$	P_N^*	$\min\{s_{N-1}, (\gamma L)^*(\kappa)\}$	$D_N(\kappa)$	Ratio (%)
1	1.939	1.135	1.500	3.994	1.500	3.994	100.000
2	1.939	1.288	1.606	7.400	1.733	7.440	100.547
5	1.939	1.882	1.748	18.310	1.893	18.456	100.798
8	1.939	2.750	1.809	29.509	1.935	29.721	100.719
9	1.939	3.121	1.823	33.254	1.939	33.483	100.689
10	1.939	3.542	1.835	36.999	1.939	37.246	100.667
20	1.939	12.546	1.885	74.170	1.939	74.867	100.940
30	1.939	44.438	1.894	111.360	1.939	112.489	101.014
40	1.939	144.899	1.898	148.645	1.939	150.111	100.986
50	1.939	182.520	1.905	186.003	1.939	187.733	100.930
100	1.939	370.629	1.916	373.255	1.939	375.841	100.693

In the weakly convex case, the sequence $\{s_k(\kappa)\}_{k=0}^{\infty}$ is truncated when it exceeds the asymptotically optimal constant stepsize for nonconvex functions from Proposition 2.15. The numerical evidence in Table 4 supports that the rate from Corollary 2.21 outperforms both: (i) optimal constant stepsize schedule $(\gamma L)^*(N,\kappa)$ (obtained by numerically maximizing the denominator from Theorem 2.4) and (ii) asymptotically optimal constant stepsize $(\gamma L)^*(\kappa)$. As κ approaches 0, $(\gamma L)^*(\kappa)$ converges to 2 and the result from Corollary 2.19 is retrieved.

3 Preliminary results for the proofs

Interpolation conditions are a set of inequalities that tightly characterize a function belonging to a specific function class $\mathcal{F}$ by focusing on a set of triplets $\mathcal{T} = \{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$. The interpolation problem itself, described in [25, 28], has its historical roots in Whitney's work on the extension of continuous functions [36, 37]. As demonstrated by Taylor et al. [32], interpolation conditions are a key tool for precise convergence analysis required to *exactly* solve the infinite-dimensional performance estimation problems (PEP) introduced in the seminal work of Drori and Teboulle [16].

Definition 3.1 ([32, Definition 2]) The set $\mathcal{T} := \{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ is called $\mathcal{F}_{\mu,L}$-interpolable if and only if there exists a function $f \in \mathcal{F}_{\mu,L}$ such that $\nabla f(x_i) = g_i$ and $f(x_i) = f_i$ for all $i \in \mathcal{I}$.

Theorem 3.2 generalizes the interpolation conditions for smooth (strongly) convex functions [32, Theorem 4] and for smooth non-necessarily functions ($\mu{=}{-}L$) [30, Theorem 3.10] to any function where $\mu{<}0$. The proof bases on standard arguments and is given in Appendix B.1. A graphical interpretation of interpolating functions for the case $\mu{=}{-}L$ is given in [29, page 71].

Theorem 3.2 *(Interpolation conditions for smooth functions) Set* $\mathcal{T}{=}\{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ *is* $\mathcal{F}_{\mu,L}$*-interpolable, where* $\mu \in (-\infty, L)$*, if and only if for every pair of indices* (i, j)*, with* $i, j \in \mathcal{I}$*, it holds:*

$$f_i{-}f_j{-}\langle g_j, x_i - x_j\rangle \geq \frac{1}{2L}\|g_i{-}g_j\|^2 + \frac{\mu}{2L(L{-}\mu)}\|g_i{-}g_j{-}L(x_i{-}x_j)\|^2. \quad (3.1)$$

Theorem 3.2 is tighter than [35, Theorem 2.2], which, to the best of our knowledge, is the first to explicitly use lower curvature information to improve stepsize ranges in weakly convex settings (the authors use it in the context of deriving convergence rates for the Douglas-Rachford splitting method).

Corollary 3.3 *Let* $f \in \mathcal{F}_{\mu,L}$*, with* $\mu \in (-\infty, L)$*. For all* $x_i, x_j \in \mathbb{R}^d$ *it holds:*

$$0 \geq \left\langle \nabla f(x_i) - \nabla f(x_j) - L(x_i - x_j)\,,\ \nabla f(x_i) - \nabla f(x_j) - \mu(x_i - x_j)\right\rangle. \quad (3.2)$$

Proof Inequality (3.2) results by summing up inequalities (3.1) written for the pairs (x_i, x_j) and (x_j, x_i) and performing simplifications. □

Corollary 3.3 is an extension to negative lower curvatures μ of the co-coercivity property from [27, Theorem 2.1.12]. When $L + \mu > 0$, the class $\mathcal{F}_{\mu,L}$ fits in the framework of $(\frac{L\mu}{L+\mu}, \frac{1}{L+\mu})$-semimonotone operators (e.g., see [17, Proposition 4.13]), which satisfy inequality (3.2) by definition.

Interpolation conditions are used in conjunction with the (GD) iterations by substituting $x_j = x_i - \sum_{k=i}^{j-1} \gamma_k \nabla f(x_k)$, as detailed in Definition 3.4. Since the proofs are limited to constant stepsizes, we constrain γ_k to γ.

Definition 3.4 Inequalities ($Q_{[i,j]}$) and ($Q_{[j,i]}$) are necessary and sufficient conditions connecting gradient descent iterations separated by $(j - i)$ steps:

$$f(x_i) - f(x_j) \geq \gamma \Big\langle \nabla f(x_j), \sum_{k=i}^{j-1} \nabla f(x_k) \Big\rangle + \frac{1}{2L} \big\| \nabla f(x_i) - \nabla f(x_j) \big\|^2 + \frac{\mu}{2L(L-\mu)} \big\| \nabla f(x_i) - \nabla f(x_j) - \gamma L \sum_{k=i}^{j-1} \nabla f(x_k) \big\|^2; \tag{$Q_{[i,j]}$}$$

$$f(x_j) - f(x_i) \geq -\gamma \Big\langle \nabla f(x_i), \sum_{k=i}^{j-1} \nabla f(x_k) \Big\rangle + \frac{1}{2L} \big\| \nabla f(x_i) - \nabla f(x_j) \big\|^2 + \frac{\mu}{2L(L-\mu)} \big\| \nabla f(x_i) - \nabla f(x_j) - \gamma L \sum_{k=i}^{j-1} \nabla f(x_k) \big\|^2. \tag{$Q_{[j,i]}$}$$

We refer to inequalities ($Q_{[i,j]}$) and ($Q_{[j,i]}$) as *distance-1* or *distance-2* inequalities if the distances between their indices satisfy $|i - j| \leq 1$ or $|i - j| \leq 2$, respectively.

In contrast to traditional convergence study, we separate the analysis concerning the optimal objective value f_* and solely focus on the function value gap up to step N, namely $f(x_0) - f(x_N)$. By making a slight modification, as detailed in the descent result from Lemma 3.5, we derive a convergence rate incorporating f_* (for a proof of the lemma, see for example [27, Section 1.2.3]).

Lemma 3.5 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, L)$. One iteration of (GD) with stepsize $\gamma L \in (0, 2)$ decreases the function values and at each point x_i it holds:*

$$f(x_i) - f_* \geq \frac{1}{2L} \|\nabla f(x_i)\|^2. \tag{3.3}$$

A convergence rate about the last gradient norm $\|\nabla f(x_N)\|$ is obtained for (strongly) convex functions, since the gradient norm is non-increasing, as shown for example in [34, Lemma 2].

Lemma 3.6 (Monotonicity of gradient norm for convex functions). *Let $f \in \mathcal{F}_{0,L}$ and one iteration with stepsize $\gamma > 0$, connecting x_i and x_{i+1}. Then:*

$$\|\nabla f(x_i)\|^2 - \|\nabla f(x_{i+1})\|^2 \geq \tfrac{2-\gamma L}{\gamma L} \|\nabla f(x_i) - \nabla f(x_{i+1})\|^2. \tag{3.4}$$

Moreover, if $\gamma \leq \frac{2}{L}$, then the gradient norm is non-increasing.

4 Proofs of performance bounds

In general, solving the associated PEP which provides (numerical) exact performance bounds in closed form is difficult. We propose a systematic procedure which eliminates the standard PEP numerical guessing by effectively canceling specific terms in a linear combination of a well defined set of inequalities from Definition 3.4 to mimic the shape of the performance metric, as described in Remark 4.10.

The PEP framework facilitated the identification and validation of worst-case scenarios, streamlining the necessary inequalities in the proofs, the subsequent demonstrations being inspired by the numerical solutions (which are not unique), with an improved understanding. Additionally, the intricate range of sublinear regimes outlined in Theorem 2.4 could only be validated numerically, but not deducible from the PEP numerical analyses. For a comprehensive understanding of the PEP setup for gradient descent, we refer the interested reader to [32] or [1].

Similar proofs using distance-1 interpolation conditions to obtain tight convergence rates on different performance metrics are given in [33, Theorem 2.1, Section 3] where the proximal gradient descent is analysed, assuming strong convexity on the smooth function in the split.

While using "distance-2" inequalities is a valid method for proving tight bounds with stepsizes greater than $\overline{\gamma L}_1(\kappa)$, we mention that the only fundamental requirement is to use conditions connecting non-consecutive iterates. In fact, PEP experiments reveal several alternative strategies, employing: both distance-2 inequalities but with different multipliers, only distance-2 inequalities of the type $Q_{[i+2,i]}$, or inequalities relating the final iterate $Q_{[N,i]}$ [23]. Our specific approach is advantageous because of its unique cancellation technique, described in Remark 4.10, which offers a methodical derivation. However, beyond this procedural benefit, we have not yet developed a deeper intuition.

Proofs organization. The lemmas from Sections 4.1 to 4.3 are employed in the rest of the section to prove the convergence rates outlined in Section 2. Table 1 provides a summary of the specific lemmas utilized in the demonstrations and the corresponding worst-case function types. The proofs are organized based on the stepsize thresholds delimiting the different regimes, even for (strongly) convex functions where the rates are expressed using minima in the denominators.

Main idea: Target inequality. Inequalities ($Q_{[i,j]}$) and ($Q_{[j,i]}$) contribute via non-negative weights $\alpha_{[i,j]}$ and $\alpha_{[j,i]}$, respectively, to a generalized inequality linking N iterations such that $\sigma_i \geq 0$:

$$\sum_{0\leq i<j\leq N} \left(\alpha_{[i,j]} Q_{[i,j]} + \alpha_{[j,i]} Q_{[j,i]}\right) : \; f(x_0) - f(x_N) \geq \sum_{i=0}^{N} \sigma_i \frac{\|\nabla f(x_i)\|^2}{2L}. \quad (\clubsuit)$$

Proposition 4.1 *(General procedure) Inequality* ($\clubsuit$) *implies:*

$$\frac{f(x_0) - f(x_N)}{\sum_{i=0}^{N} \sigma_i} \geq \frac{1}{2L} \min_{0\leq i\leq N} \left\{\|\nabla f(x_i)\|^2\right\},$$

with equality only if all gradient norms $\|\nabla f(x_i)\|^2$ *with* $\sigma_i > 0$ *are equal. Moreover,*

$$\frac{f(x_0)-f_*}{1+\sum_{i=0}^N \sigma_i} \geq \frac{1}{2L} \min_{0\leq i\leq N} \left\{\|\nabla f(x_i)\|^2\right\}.$$

Proof The first inequality results after taking minimum gradient norm in (♣). The adjustment in the denominator of the second inequality results from combining (3.3) with (♣) and taking again the minimum gradient norm. □

Following Proposition 4.1, all convergence proofs reduce to demonstrating an inequality of the form (♣) with positive weights σ_i. We utilize interpolation inequalities ($Q_{[i,j]}$) and ($Q_{[j,i]}$) with (i) $|j-i|=1$ to prove Theorem 2.5 and (ii) $|j-i|\in\{1,2\}$ to prove Theorem 2.4 and Theorem 2.3.

Notation. For readability reasons, we sometimes denote $f_i=f(x_i)$ and $g_i=\nabla f(x_i)$. Recall: $\kappa=\frac{\mu}{L}$ is the curvature ratio; $\gamma L \in (0,2)$ is the (normalized) stepsize and $\overline{\gamma L}_\infty(\kappa)=\frac{2}{1+[\kappa]_+}$.

4.1 Distance-1 lemmas for weakly convex functions

For $|j-i|=1$, inequalities ($Q_{[i,j]}$) and ($Q_{[j,i]}$) simplify to:

$$f_i - f_{i+1} \geq \frac{\gamma\mu\,\gamma L - 2\gamma\mu + 1}{L-\mu}\frac{\|g_i\|^2}{2} + \frac{1}{L-\mu}\frac{\|g_{i+1}\|^2}{2} + \frac{\gamma L-1}{L-\mu}\langle g_i, g_{i+1}\rangle; \tag{$Q_{[i,i+1]}$}$$

$$f_{i+1} - f_i \geq \frac{\gamma\mu\gamma L - 2\gamma L + 1}{L-\mu}\frac{\|g_i\|^2}{2} + \frac{1}{L-\mu}\frac{\|g_{i+1}\|^2}{2} + \frac{\gamma\mu-1}{L-\mu}\langle g_i, g_{i+1}\rangle. \tag{$Q_{[i+1,i]}$}$$

Note that their l.h.s. only contains consecutive function values, whereas the r.h.s. is a combination of gradients, with symmetric coefficients in γL and $\gamma\mu$.

Lemma 4.2 *Let* $f\in\mathcal{F}_{\mu,L}$, *with* $\mu\in(-\infty,L)$, *and consider one iteration of* (GD) *with stepsize* $\gamma L\in(0,1]$. *Then the following inequality holds, with equality only if* $\nabla f(x_i)=\nabla f(x_{i+1})$ *or* $\gamma L=1$*:*

$$f(x_i)-f(x_{i+1}) \geq \frac{\gamma L\,\gamma\mu - 2\gamma\mu + \gamma L}{L-\mu}\frac{\|\nabla f(x_i)\|^2}{2} + \frac{\gamma L}{L-\mu}\frac{\|\nabla f(x_{i+1})\|^2}{2}. \tag{N2SD}$$

Furthermore, if $\mu\leq 0$, *then the gradient norms from* (N2SD) *are weighted by positive quantities.*

Proof Inequality ($Q_{[i,i+1]}$) is reformulated by expressing the inner product in terms of gradient norms, i.e., $2\langle g_i, g_{i+1}\rangle = \|g_i\|^2 + \|g_{i+1}\|^2 - \|g_i - g_{i+1}\|^2$, as:

$$f_i - f_{i+1} \geq \frac{\gamma L\, \gamma\mu - 2\gamma\mu + \gamma L}{L-\mu} \frac{\|g_i\|^2}{2} + \frac{\gamma L}{L-\mu} \frac{\|g_{i+1}\|^2}{2} + \frac{1-\gamma L}{L-\mu} \frac{\|g_i - g_{i+1}\|^2}{2}. \tag{4.1}$$

Since $\gamma L \in (0, 1]$, the mixed term can be neglected, resulting in inequality (N2SD), with equality only if $g_i = g_{i+1}$ or $\gamma L = 1$. The numerator of first term rewrites as $-\gamma\mu(1 - \gamma L) + (\gamma L - \gamma\mu)$, hence it is positive when $\mu \leq 0$. □

Lemma 4.2 does not hold for stepsizes $\gamma L \in (1, 2)$.

Lemma 4.3 *Let $f \in \mathcal{F}_{\mu,L}$, with, $\mu \in (-\infty, L)$, and consider one iteration of (GD) with stepsize $\gamma L \in [1, \overline{\gamma L}_\infty(\kappa))$. Then the following inequality holds:*

$$f(x_i) - f(x_{i+1}) \geq \frac{\gamma[(2-\gamma L)(2-\gamma\mu)-1]}{2-\gamma L-\gamma\mu} \frac{\|\nabla f(x_i)\|^2}{2} + \frac{\gamma}{2-\gamma L-\gamma\mu} \frac{\|\nabla f(x_{i+1})\|^2}{2}. \tag{N4SD}$$

When $\gamma L > 1$, equality holds only if:

$$\langle \nabla f(x_i), \nabla f(x_{i+1})\rangle = \frac{(1-\gamma\mu)(1-\gamma L)}{2-\gamma L-\gamma\mu} \|\nabla f(x_i)\|^2 + \frac{1}{2-\gamma L-\gamma\mu} \|\nabla f(x_{i+1})\|^2. \tag{4.2}$$

Furthermore, if $\gamma L \in [1, \overline{\gamma L}_1(\kappa)]$, then the gradient norms from (N4SD) are weighted by nonnegative quantities.

Proof The mixed term in inequality (4.1) enters with a negative contribution for $\gamma L > 1$. To counterbalance it, the complementary interpolation inequality ($Q_{[i+1,i]}$) is used, written with the equivalent expression:

$$f_{i+1} - f_i \geq \frac{\gamma L\gamma\mu - 2\gamma L + \gamma\mu}{L-\mu} \frac{\|g_i\|^2}{2} + \frac{\gamma\mu}{L-\mu} \frac{\|g_{i+1}\|^2}{2} + \frac{1-\gamma\mu}{L-\mu} \frac{\|g_i - g_{i+1}\|^2}{2}.$$

Consider the sum of distance-1 inequalities $\bar{I}_i^1 := Q_{[i,i+1]} + Q_{[i+1,i]}$:

$$0 \geq \frac{2\gamma L\gamma\mu - \gamma L - \gamma\mu}{L-\mu} \frac{\|g_i\|^2}{2} + \frac{\gamma L + \gamma\mu}{L-\mu} \frac{\|g_{i+1}\|^2}{2} + \frac{2-\gamma L-\gamma\mu}{L-\mu} \frac{\|g_i - g_{i+1}\|^2}{2}. \tag{$\bar{I}_i^1$}$$

By weighting ($\bar{I}_i^1$) with $\beta(\gamma L, \gamma\mu) := \frac{\gamma L - 1}{2-\gamma L-\gamma\mu} \geq 0$ and summing it with (4.1), the coefficient of $\|g_i - g_{i+1}\|^2$ gets canceled and we obtain (N4SD). The equality condition (4.2) follows from the equality case of ($\bar{I}_i^1$). Nevertheless, the coefficient of $\|g_i\|^2$ is nonnegative for $\gamma L \in [1, \overline{\gamma L}_1(\kappa)]$, where $\overline{\gamma L}_1(\kappa)$ is the largest stepsize $\gamma L \in [1, 2)$

for which: $(2-\gamma L)(2-\gamma\mu)-1 \geq 0$, implying nonnegativity of the weight of $\|\nabla f(x_i)\|^2$ in (N4SD). □

The threshold $\overline{\gamma L}_1(\kappa)$ has the explicit formula $\overline{\gamma L}_1(\kappa) = \frac{3}{1+\kappa+\sqrt{1-\kappa+\kappa^2}}$.

Remark 4.4 (Sufficient decrease results) Inequalities (N2SD) and (N4SD) are extensions to *weakly convex* functions of the "double" sufficient decrease formula (2SD) [34, Lemma 1] and the "quadruple sufficient decrease" formula (4SD) [34, Lemma 5], respectively, obtained for convex functions ($\mu = 0$):

$$f(x_i)-f(x_{i+1}) \geq \gamma\frac{\|\nabla f(x_i)\|^2}{2} + \gamma\frac{\|\nabla f(x_{i+1})\|^2}{2}, \quad \forall \gamma L \in (0,1]; \tag{2SD}$$

$$f(x_i)-f(x_{i+1}) \geq \frac{\gamma(3-2\gamma L)}{2-\gamma L}\frac{\|\nabla f(x_i)\|^2}{2} + \frac{\gamma}{2-\gamma L}\frac{\|\nabla f(x_{i+1})\|^2}{2}, \quad \forall \gamma L \in [1,2). \tag{4SD}$$

4.2 Distance-1 lemmas for strongly convex functions

Further on, the analysis focuses on constant stepsizes $\gamma L \in (1,2)$, interval only partially covered by Theorem 2.5.

Additional notation. For brevity, we sometimes use $\rho{=}\rho(\gamma L){:=}1{-}\gamma L \in (-1,0)$, $\eta{=}\eta(\gamma\mu){:=}1{-}\gamma\mu$. Then $T_k(\gamma L,\kappa) = E_k(1{-}\kappa\gamma L){-}E_k(1{-}\gamma L)$ (see Definition 2.9), where $E_k(x) = \sum_{j=1}^{2k} x^{-j}$, can be equivalently written as $T_k(\rho,\eta) := T_k(\gamma L,\kappa)$, with $T_k(\rho,\eta) = E_k(\eta) - E_k(\rho)$. Recall $\overline{\gamma L}_\infty(\kappa) = \frac{2}{1+[\kappa]_+}$.

Definition 4.5 Let $\bar{N}(\gamma L,\kappa) := \max\{k \in \mathbb{N} \mid T_k(\gamma L,\kappa) \geq 0\}$.

Proposition 4.6 *Let $\mu \in (-\infty, L)$ and $\gamma L \in \left(1, \overline{\gamma L}_\infty(\kappa)\right)$. Then $\eta > 0$, $\eta+\rho > 0$ and $\rho^{-2} > \eta^{-2}$.*

Proof Condition $\gamma L < \overline{\gamma L}_\infty(\kappa)$ implies $\gamma L < \frac{2}{1+[\kappa]_+}$, hence $0 < 2{-}\gamma L{-}\gamma L[\kappa]_+ \leq 2{-}\gamma L{-}\gamma L\kappa$, implying $\eta+\rho > 0$. Since $\rho = 1-\gamma L \in (-1,0)$, we have $\eta = 1-\gamma\mu > 0$. Condition $L > \mu$ implies $\eta > \rho$ and together with $\eta+\rho > 0$ we conclude that $\eta^2 > \rho^2$, therefore $\rho^{-2} > \eta^{-2}$. □

Lemma 4.7 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, L)$, and consider one iteration of (GD) from x_i with stepsize $\gamma L \in (1,2)$. Then it holds:*

$$\begin{aligned}\frac{f(x_i)-f(x_{i+1})}{\gamma} \geq -E_i(\rho)\frac{\|\nabla f(x_i)\|^2}{2} + E_{i+1}(\rho)\frac{\|\nabla f(x_{i+1})\|^2}{2} + \\ + \frac{1-(\eta+\rho)E_{i+1}(\rho)}{\eta-\rho}\frac{\|\nabla f(x_{i+1})-\rho\nabla f(x_i)\|^2}{2}.\end{aligned} \tag{4.3}$$

Furthermore, if $\mu \in (0,L)$ and $\gamma L \in [\frac{2}{1+\kappa}, 2)$, then $1-(\eta+\rho)E_i(\rho) > 0$.

Proof Let $\beta_{i+1}(\rho) := (-\rho)E_{i+1}(\rho) > 0$. Inequality (4.3) results after performing the simplifications in the linear combination of interpolation inequalities:

$$Q_{[i,i+1]} + \beta_{i+1}(\rho) \times (Q_{[i,i+1]} + Q_{[i+1,i]}),$$

where $(Q_{[i,j]})$ and $(Q_{[j,i]})$ are equivalently rewritten for the pair $(i, i+1)$ as:

$$\begin{aligned}(1+\beta_{i+1})\times Q_{[i,i+1]}: \ & \frac{f_i - f_{i+1}}{\gamma} \geq (1+\rho)\frac{\|g_i\|^2}{2} + \frac{1}{\eta-\rho}\frac{\|g_{i+1}-\rho g_i\|^2}{2};\\ \beta_{i+1}\times Q_{[i+1,i]}: \ & \frac{f_{i+1} - f_i}{\gamma} \geq -\frac{\|g_i\|^2}{2} - \frac{1}{\rho}\frac{\|g_{i+1}\|^2}{2} + \frac{\eta}{\rho}\frac{1}{\eta-\rho}\frac{\|g_{i+1}-\rho g_i\|^2}{2}.\end{aligned}$$

With $\gamma L \in [\frac{2}{1+\kappa}, 2)$, we have $\eta+\rho \leq 0$ (see Proposition 4.6) and therefore $1-(\eta+\rho)E_{i+1}(\rho) > 0$. □

Lemma 4.8 *Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, L)$, and consider one iteration of (GD) from x_i with stepsize $\gamma L \in \left(0, \overline{\gamma L}_{\infty}(\kappa)\right)$. Then it holds:*

$$\begin{aligned}\frac{f(x_i) - f(x_{i+1})}{\gamma} \geq & -E_i(\eta)\frac{\|\nabla f(x_i)\|^2}{2} + E_{i+1}(\eta)\frac{\|\nabla f(x_{i+1})\|^2}{2} + \\ & + \frac{-1+(\eta+\rho)E_{i+1}(\eta)}{\eta-\rho}\frac{\|\nabla f(x_{i+1}) - \eta\nabla f(x_i)\|^2}{2}.\end{aligned} \tag{4.4}$$

Furthermore, if $\mu \in (0, L)$ and $\gamma L \in \left(0, \overline{\gamma L}_{i+1}(\kappa)\right)$, then $-1+(\eta+\rho)E_{i+1}(\eta) > 0$.

Proof Firstly, note that $\eta = 1-\gamma\mu > 0$; the case $\gamma L \in \left(1, \frac{2}{1+[\kappa]_+}\right)$ is covered by Proposition 4.6, while for $\gamma L \in (0, 1]$ we have $1-\gamma\mu > 1-\gamma L \geq 0$. Let:

$$\beta_{i+1}(\eta) := -1 + \eta E_{i+1}(\eta) = \frac{1+\eta+\cdots+\eta^{2i}}{\eta^{2i+1}} > 0.$$

Inequality (4.4) results after performing the simplifications in the linear combination of interpolation inequalities:

$$Q_{[i,i+1]} + \beta_{i+1}(\eta) \times (Q_{[i,i+1]} + Q_{[i+1,i]}),$$

where $(Q_{[i,j]})$ and $(Q_{[j,i]})$ equivalently rewrite for the pair $(i, i+1)$ as:

$$\begin{aligned}(1+\beta_{i+1})\times Q_{[i,i+1]}: \ & \frac{f_i - f_{i+1}}{\gamma} \geq \frac{\|g_i\|^2}{2} + \frac{1}{\eta}\frac{\|g_{i+1}\|^2}{2} + \frac{\frac{\rho}{\eta}}{\eta-\rho}\frac{\|g_{i+1}-\eta g_i\|^2}{2};\\ \beta_{i+1}\times Q_{[i+1,i]}: \ & \frac{f_{i+1} - f_i}{\gamma} \geq -(1+\eta)\frac{\|g_i\|^2}{2} + \frac{1}{\eta-\rho}\frac{\|g_{i+1}-\eta g_i\|^2}{2}.\end{aligned}$$

Condition $\gamma L < \overline{\gamma L}_{i+1}(\kappa)$ implies $T_{i+1}(\rho,\eta) < 0$, hence $i \geq \bar{N}(\gamma L, \kappa)$ (recall Definition 4.5: $T_{\bar{N}}(\rho,\eta) \geq 0$ and $T_{\bar{N}+1}(\rho,\eta) < 0$, corresponding to $\gamma L < \overline{\gamma L}_{\bar{N}+1}(\kappa)$). Then the coefficient of the mixed term in (4.4) is positive:

$$\frac{-1+(\eta+\rho)E_{i+1}(\eta)}{\eta-\rho} \geq \frac{-1+(\eta+\rho)E_{\bar{N}+1}(\eta)}{\eta-\rho} = \frac{T_{\bar{N}}(\rho,\eta)-\rho^2 T_{\bar{N}+1}(\rho,\eta)}{(\eta-\rho)^2} > 0.$$

The first inequality results from the monotonic increase with index i of $E_{i+1}(\eta)$, whereas the last one by the definition of $\bar{N}(\gamma L, \kappa)$. □

4.3 Distance-2 lemmas

In this section we derive descent lemmas involving distance-2 interpolation inequalities, needed to prove rates on the stepsize range $\gamma L \in (1, \overline{\gamma L}_{\infty}(\kappa))$.

Lemma 4.9 *(Central distance-2 result) Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, L)$. Consider $k+1$ iterations of (GD) with constant stepsize $\gamma L \in \left[\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{\infty}(\kappa)\right)$ starting from x_0, where $k \geq 1$. Then it holds:*

$$\frac{f(x_0)-f(x_k)}{\gamma} + \frac{(\eta+\rho)T_k(\rho,\eta)}{2(\eta-\rho)}\frac{f(x_k)-f(x_{k+1})}{\gamma} \geq (E_k(\eta)+E_k(\rho))\frac{\|\nabla f(x_k)\|^2}{4} + \\ + \frac{T_k(\rho,\eta)}{2(\eta-\rho)}\Big[\eta\rho\frac{\|\nabla f(x_k)\|^2}{2} + \langle \nabla f(x_k), \nabla f(x_{k+1})\rangle + \frac{\|\nabla f(x_{k+1})\|^2}{2}\Big]. \qquad \text{(D2)}$$

Proof Consider two iterations connecting x_i, x_{i+1} and x_{i+2}, with $i \geq 0$. The interpolation inequalities ($Q_{[i,j]}$) and ($Q_{[j,i]}$) written for distance-1 ($Q_{[i,i+1]}$ and $Q_{[i+1,i]}$) and distance-2 ($Q_{[i,i+2]}$ and $Q_{[i+2,i]}$), respectively:

$$Q_{[i,i+1]}\colon \frac{f_i - f_{i+1}}{\gamma} \geq (1+\rho)\frac{\|g_i\|^2}{2} + \frac{1}{\eta-\rho}\frac{\|g_{i+1}-\rho g_i\|^2}{2};$$

$$Q_{[i+1,i]}\colon \frac{f_{i+1} - f_i}{\gamma} \geq -\frac{\|g_i\|^2}{2} - \frac{1}{\rho}\frac{\|g_{i+1}\|^2}{2} + \frac{\eta}{\rho}\frac{1}{\eta-\rho}\frac{\|g_{i+1}-\rho g_i\|^2}{2};$$

$$Q_{[i,i+2]}\colon \frac{f_i - f_{i+2}}{\gamma} \geq (1+\rho+\rho^2)\frac{\|g_i\|^2}{2} + \rho\frac{\|g_{i+1}\|^2}{2} + \frac{1}{\rho(\eta-\rho)}\times \\ \Big[\rho(1-\eta)\frac{\|g_{i+1}-\rho g_i\|^2}{2} - (1-\rho)\frac{\|g_{i+2}-\rho g_{i+1}\|^2}{2} + \frac{\|g_{i+2}-\rho^2 g_i\|^2}{2}\Big];$$

$$Q_{[i+2,i]}\colon \frac{f_{i+2} - f_i}{\gamma} \geq -\frac{\|g_i\|^2}{2} - \frac{1}{\rho}\frac{\|g_{i+1}\|^2}{2} - \frac{1+\rho}{\rho^2}\frac{\|g_{i+2}\|^2}{2} + \frac{1}{\rho(\eta-\rho)}\times \\ \Big[\eta(1-\rho)\frac{\|g_{i+1}-\rho g_i\|^2}{2} - (1-\eta)\frac{\|g_{i+2}-\rho g_{i+1}\|^2}{2} + \frac{\eta}{\rho}\frac{\|g_{i+2}-\rho^2 g_i\|^2}{2}\Big],$$

are summed up in a linear combination with *nonnegative* corresponding weights:

$$(4.5) \equiv \alpha_{[i,i+1]} Q_{[i,i+1]} + \alpha_{[i+1,i]} Q_{[i+1,i]} + \alpha_{[i,i+2]} Q_{[i,i+2]} + \alpha_{[i+2,i]} Q_{[i+2,i]},$$

where:

$$\begin{aligned}
\alpha_{[i,i+1]} &:= 1 - \rho E_{i+1}(\rho) + \frac{(1-\eta)T_i(\rho,\eta)}{2(\eta-\rho)} - \frac{\eta(1+\rho)T_{i+1}(\rho,\eta)}{2(\eta-\rho)};\\
\alpha_{[i+1,i]} &:= -\rho E_{i+1}(\rho) + \frac{(1+\rho)T_i(\rho,\eta)}{2(\eta-\rho)} + \frac{\rho(1-\eta)T_{i+1}(\rho,\eta)}{2(\eta-\rho)};\\
\alpha_{[i,i+2]} &:= \frac{\eta T_{i+1}(\rho,\eta)}{2(\eta-\rho)};\\
\alpha_{[i+2,i]} &:= \frac{-\rho T_{i+1}(\rho,\eta)}{2(\eta-\rho)},
\end{aligned}$$

and (4.5) is obtained after performing the simplifications:

$$\begin{aligned}
&\tfrac{f_i - f_{i+1}}{\gamma} + (\eta+\rho)\Big[-\tfrac{T_i(\rho,\eta)}{2(\eta-\rho)}\tfrac{f_i - f_{i+1}}{\gamma} + \tfrac{T_{i+1}(\rho,\eta)}{2(\eta-\rho)}\tfrac{f_{i+1} - f_{i+2}}{\gamma}\Big] \geq\\
&-\tfrac{T_i(\rho,\eta)}{2(\eta-\rho)}\Big[\eta\rho\tfrac{\|g_i\|^2}{2} + \langle g_i, g_{i+1}\rangle + \tfrac{\|g_{i+1}\|^2}{2}\Big] - (E_i(\eta)+E_i(\rho))\tfrac{\|g_i\|^2}{4}\\
&+\tfrac{T_{i+1}(\rho,\eta)}{2(\eta-\rho)}\Big[\eta\rho\tfrac{\|g_{i+1}\|^2}{2} + \langle g_{i+1}, g_{i+2}\rangle + \tfrac{\|g_{i+2}\|^2}{2}\Big] + \big(E_{i+1}(\eta)+E_{i+1}(\rho)\big)\tfrac{\|g_{i+1}\|^2}{4}.
\end{aligned} \tag{4.5}$$

Telescoping it for indices $i = 0, 1, \ldots, k-1$ yields inequality (D2). It remains to show the nonnegativity of multipliers. Since the stepsize thresholds increase with index i, condition $\gamma L \geq \overline{\gamma L}_k(\kappa)$ ensures $T_{i+1}(\rho,\eta) \geq 0$, for all $i = 0, \ldots, k-1$. This directly implies the positivity of $\alpha_{[i,i+2]}$ and $\alpha_{[i+2,i]}$. The positivity of distance-1 multipliers is evidenced by their equivalent representations as nonnegative sums (recall $\rho \in (-1, 0)$, $\eta \in (0, \infty)$; $\eta + \rho > 0$ and $\rho^{-2} > \eta^{-2}$ from Proposition 4.6):

$$\begin{aligned}
2\alpha_{[i,i+1]} &= 1 + \frac{\eta + \eta^{-2i}}{1+\eta} + \frac{(1+\rho)(\eta+\rho-\rho\eta)}{\eta\rho^2}\frac{-1+\rho^{-2(i+1)}}{-1+\rho^{-2}} +\\
&\qquad \frac{(1+\rho)(1+\eta^2)}{\eta(\eta-\rho)}\sum_{j=0}^{i}\big(\rho^{-2j} - \eta^{-2j}\big) > 0;\\
2\alpha_{[i+1,i]} &= \frac{-1+\rho^{-2i}}{1-\rho} + \frac{\eta(1+\rho)}{1-\rho}\frac{-1+\eta^{-2(i+1)}}{1-\eta} +\\
&\qquad + \frac{-\rho\big[1-\rho+\eta(1+\rho)\big]}{1-\rho}\frac{\rho^{-2(i+1)} - \eta^{-2(i+1)}}{\eta-\rho} > 0.
\end{aligned}$$

□

Remark 4.10 (Derivation of Lemma 4.9) Inequality (D2) selectively involves gradient norms of indices $\{k, k+1\}$, function values of indices $\{0, k, k+1\}$ and the inner product

$\langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle$. As a result, the linear combination of interpolation inequalities eliminates all other function values and gradients. Specifically, it removes: k gradient norms, k consecutive and k distance-2 inner products of gradients, $k-1$ function values, while setting the coefficient of f_0 to $\frac{1}{\gamma}$, totaling $4k$ constraints. Remarkably, the linear combination employs exactly $4k$ inequalities as well, highlighting why we needed to involve distance-2 interpolation inequalities. Consequently, deriving Lemma 4.9 (and subsequent convergence rates) entails solving a linear system, contrasting with the PEP numerical analysis, which involves solving an SDP.

Subsequent lemmas involving distance-2 interpolation inequalities are derived by augmenting Lemma 4.9 with distance-1 interpolation inequalities.

Lemma 4.11 *(Multistep descent) Let $f \in \mathcal{F}_{\mu,L}$, with $\mu \in (-\infty, L)$, and consider $k+1 \geq 1$ iterations of (GD) with stepsize $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_\infty(\kappa))$, satisfying $\gamma L > 1$, starting from x_0. Then the following inequality holds:*

$$\frac{f(x_0)-f(x_{k+1})}{\gamma} \geq \frac{-\eta^2\rho^2 T_{k+1}(\rho,\eta)}{\eta^2-\rho^2}\frac{\|\nabla f(x_k)\|^2}{2} + \frac{T_k(\rho,\eta)+(\eta-\rho)}{\eta^2-\rho^2}\frac{\|\nabla f(x_{k+1})\|^2}{2}. \quad \text{(GN4SD)}$$

Furthermore, if $\mu \in (-\infty, 0]$ and $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa))$, then the gradient norms from (GN4SD) are weighted by nonnegative quantities.

Proof For $k = 0$, the result follows from Lemma 4.3 and we obtain exactly (N4SD). For $k \geq 1$, inequality (GN4SD) is derived by adjusting (D2) through the linear combination of interpolation inequalities $\alpha_{[k,k+1]} Q_{[k,k+1]} + \alpha_{[k+1,k]} Q_{[k+1,k]}$:

$$\begin{aligned}
&\left[1 - \frac{(\eta+\rho)\,T_k(\rho,\eta)}{2(\eta-\rho)}\right]\frac{f_k - f_{k+1}}{\gamma} \\
&\quad \geq \left(- E_k(\eta) - E_k(\rho) - \frac{\eta\rho T_k(\rho,\eta)}{\eta-\rho} - \frac{2\eta^2\rho^2 T_{k+1}(\rho,\eta)}{\eta^2-\rho^2}\right)\frac{\|g_k\|^2}{4} \\
&\quad + \left(\frac{\eta-\rho+T_k(\rho,\eta)}{\eta^2-\rho^2} - \frac{T_k(\rho,\eta)}{2(\eta-\rho)}\right)\frac{\|g_{k+1}\|^2}{2} - \frac{T_k(\rho,\eta)}{2(\eta-\rho)}\langle g_k, g_{k+1}\rangle,
\end{aligned}$$

where the corresponding weights are defined as follows:

$$\begin{aligned}
\alpha_{[k,k+1]} &:= 1 + \frac{-\rho}{\eta+\rho} + \frac{(1-\eta)(\eta+\rho)-2\rho}{2(\eta^2-\rho^2)}T_k(\rho,\eta); \\
\alpha_{[k+1,k]} &:= \frac{-\rho}{\eta+\rho} + \frac{(1+\rho)(\eta+\rho)-2\rho}{2(\eta^2-\rho^2)}T_k(\rho,\eta).
\end{aligned}$$

Multiplier $\alpha_{[k+1,k]}$ is represented as a sum of positive quantities ($\rho < 0$). Multiplier $\alpha_{[k,k+1]}$ is directly written as sum of positive quantities if $(1-\eta)(\eta+\rho)-2\rho > 0$ (in particular, this covers the (strongly) convex case where $\eta \leq 1$). Otherwise, $\alpha_{[k,k+1]}$ is equivalently expressed as a sum of positive terms as follows, with $\eta > 1$:

$$\alpha_{[k,k+1]} = \frac{\eta}{2(\eta+\rho)} + \frac{1}{2(\eta^2-\rho^2)}\left[\frac{-\rho(\eta-\rho)(1+\eta^2)}{(\eta-1)(1-\rho)} + \right.$$

$$-\Big[(1-\eta)(\eta+\rho)-2\rho\Big]\Big(\frac{\eta^{-2k}}{\eta-1}+\frac{\rho^{-2k}}{1-\rho}\Big)\Big]>0.$$

□

Lemma 4.12 *Let $f\in\mathcal{F}_{\mu,L}$, with $\mu\in(-\infty,L)$, and consider $k+1\geq 1$ iterations of (GD) with stepsize $\gamma L\in[\overline{\gamma L}_k(\kappa),\overline{\gamma L}_\infty(\kappa))$, satisfying $\gamma L>1$, starting from x_0. Then the following inequality holds:*

$$\frac{f(x_0)-f(x_{k+1})}{\gamma}\geq E_{k+1}(\rho)\frac{\|\nabla f(x_{k+1})\|^2}{2}+\frac{\eta^2T_{k+1}(\rho,\eta)}{(\eta-\rho)^2}\frac{\|\nabla f(x_{k+1})-\rho\nabla f(x_k)\|^2}{2}. \tag{4.6}$$

Furthermore, if $\gamma L\in[\overline{\gamma L}_{k+1}(\kappa),\overline{\gamma L}_\infty(\kappa))$, then $T_{k+1}(\rho,\eta)\geq 0$.

Proof For $k=0$, the result follows from Lemma 4.7. For $k\geq 1$, inequality (4.6) is derived by adjusting (D2) with the linear combination of interpolation inequalities $\alpha_{[k,k+1]}Q_{[k,k+1]}+\alpha_{[k,k+1]}Q_{[k+1,k]}$:

$$\begin{aligned}\Big[1-\frac{(\eta+\rho)\,T_k(\rho,\eta)}{2(\eta-\rho)}\Big]\frac{f_k-f_{k+1}}{\gamma}\geq\Big(-\frac{T_k(\rho,\eta)}{2(\eta-\rho)}-\frac{\eta^2\rho T_{k+1}(\rho,\eta)}{(\eta-\rho)^2}\Big)\langle g_k,g_{k+1}\rangle+\\+\Big[\frac{2\eta^2\rho^2T_{k+1}(\rho,\eta)}{(\eta-\rho)^2}-E_k(\eta)-E_k(\rho)-\frac{\eta\rho T_k(\rho,\eta)}{\eta-\rho}\Big]\frac{\|g_k\|^2}{4}+\\+\Big[\frac{\eta^2\,T_{k+1}(\rho,\eta)}{(\eta-\rho)^2}+E_{k+1}(\rho)-\frac{T_k(\rho,\eta)}{2(\eta-\rho)}\Big]\frac{\|g_{k+1}\|^2}{2},\end{aligned}$$

where the corresponding weights are defined as:

$$\begin{aligned}\alpha_{[k,k+1]}&:=1+(-\rho)E_{k+1}(\rho)+\frac{1-\eta}{2(\eta-\rho)}T_k(\rho,\eta);\\\alpha_{[k+1,k]}&:=(-\rho)E_{k+1}(\rho)+\frac{1+\rho}{2(\eta-\rho)}T_k(\rho,\eta).\end{aligned}$$

Multiplier $\alpha_{[k+1,k]}$ is represented as a sum of positive quantities, while $\alpha_{[k,k+1]}$ is directly written as sum of positive terms if $\eta\leq 1$ (the (strongly) convex case). Otherwise, $\eta>1$ and $\alpha_{[k,k+1]}$ is equivalently expressed as sum of positive terms as:

$$\alpha_{[k,k+1]}=\frac{1+(-\rho)\rho^{-2(k+1)}}{2(1-\rho)}+\eta\frac{\eta\eta^{-2(k+1)}+(-\rho)\rho^{-2(k+1)}}{2(\eta-\rho)}.$$

□

Lemma 4.13 *Let $f\in\mathcal{F}_{\mu,L}$, with $\mu\in(-\infty,L)$, and consider $k+1\geq 1$ iterations of (GD) with stepsize $\gamma L\in[\overline{\gamma L}_k(\kappa),\overline{\gamma L}_\infty(\kappa))$, satisfying $\gamma L>1$, starting from x_0. Then the following inequality holds:*

$$\frac{f(x_0)-f(x_{k+1})}{\gamma}\geq E_{k+1}(\eta)\frac{\|\nabla f(x_{k+1})\|^2}{2}+\frac{-\rho^2T_{k+1}(\rho,\eta)}{(\eta-\rho)^2}\frac{\|\nabla f(x_{k+1})-\eta\nabla f(x_k)\|^2}{2}. \tag{4.7}$$

Furthermore, if $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa))$, *then* $T_{k+1}(\rho,\eta) < 0$.

Proof Case $k = 0$ results from Lemma 4.9; further on, $k \geq 1$. Inequality (3.2) written for the pair $(i, j) = (k, k+1)$ reads:

$$0 \geq \frac{\eta\rho}{\eta-\rho}\|g_k\|^2 + \frac{1}{\eta-\rho}\|g_{k+1}\|^2 - \frac{\eta+\rho}{\eta-\rho}\langle g_k\,,\, g_{k+1}\rangle. \tag{4.8}$$

1. **Case** $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa))$**.** By using $\eta > 0$, (4.8) equivalently rewrites as:

$$0 \geq -\eta^2\frac{\|g_k\|^2}{2} + \frac{\|g_{k+1}\|^2}{2} + \frac{\eta+\rho}{\eta-\rho}\frac{\|g_{k+1}-\eta g_k\|^2}{2}.$$

By multiplying it with $\frac{-\rho^2}{\eta^2-\rho^2}T_{k+1}(\rho,\eta)$ (positive because $T_{k+1}(\rho,\eta) < 0$) and adding it to (GN4SD), the coefficient of $\|g_k\|^2$ is canceled and we get:

$$\frac{f_0-f_{k+1}}{\gamma} \geq \frac{T_k(\rho,\eta)+(\eta-\rho)-\rho^2T_{k+1}(\rho,\eta)}{\eta^2-\rho^2}\frac{\|g_{k+1}\|^2}{2} + \frac{-\rho^2T_{k+1}(\rho,\eta)}{(\eta-\rho)^2}\frac{\|g_{k+1}-\eta g_k\|^2}{2},$$

which after simplifications is exactly (4.7).

2. **Case** $\gamma L \in \left[\overline{\gamma L}_{k+1}(\kappa), \overline{\gamma L}_{\infty}(\kappa)\right)$**.** After multiplying (4.8) with $\frac{-\rho\eta}{\eta-\rho} > 0$:

$$0 \geq \frac{\|g_{k+1}\|^2}{2} + \frac{-\eta^2}{(\eta-\rho)^2}\frac{\|g_{k+1}-\rho g_k\|^2}{2} + \frac{-\rho^2}{(\eta-\rho)^2}\frac{\|g_{k+1}-\eta g_k\|^2}{2}.$$

By scaling it with $T_{k+1}(\rho,\eta) > 0$ and summing it with (4.6), the coefficient of $\|g_{k+1}-\rho g_k\|^2$ is canceled and after simplifications we get (4.7). □

4.4 Proof of performance bounds for weakly convex functions

Proof (**Theorem 2.5**) By telescoping the sufficient decrease inequalities (N2SD) and (N4SD) for indices $i = 1, \ldots, N$ and stepsizes $\gamma_0, \ldots, \gamma_{N-1}$, respectively:

$$\begin{aligned} f_0-f_N \geq{}& \frac{\|g_0\|^2}{2}\Big[\frac{\gamma_0 L\,\gamma_0\mu-2\gamma_0\mu+\gamma_0 L}{L-\mu}\delta(\gamma_0\leq\tfrac{1}{L}) + \frac{\gamma_0[(2-\gamma_0 L)(2-\gamma_0\mu)-1]}{2-\gamma_0 L-\gamma_0\mu}\delta(\gamma_0>\tfrac{1}{L})\Big] + \\ &+ \sum_{i=1}^{N-1}\frac{\|g_i\|^2}{2}\Big[\frac{\gamma_{i-1}}{L-\mu}\delta(\gamma_{i-1}\leq\tfrac{1}{L}) + \frac{\gamma_{i-1}}{2-\gamma_{i-1}L-\gamma_{i-1}\mu}\delta(\gamma_{i-1}>\tfrac{1}{L}) + \\ &\quad + \frac{\gamma_i L\,\gamma_i\mu-2\gamma_i\mu+\gamma_i L}{L-\mu}\delta(\gamma_i\leq\tfrac{1}{L}) + \frac{\gamma_i[(2-\gamma_i L)(2-\gamma_i\mu)-1]}{2-\gamma_i L-\gamma_i\mu}\delta(\gamma_i>\tfrac{1}{L})\Big] \\ &+ \frac{\|g_N\|^2}{2}\Big[\frac{\gamma_{N-1}}{L-\mu}\delta(\gamma_{N-1}\leq\tfrac{1}{L}) + \frac{\gamma_{N-1}}{2-\gamma_{N-1}L-\gamma_{N-1}\mu}\delta(\gamma_{N-1}>\tfrac{1}{L})\Big], \end{aligned} \tag{4.9}$$

where $\delta : x \mapsto \{0, 1\}$ is the indicator function (it is 1 for true argument and 0 otherwise). The telescoped inequality resembles a decrease inequality of type (♣) if all gradient norms have positive weights; a sufficient condition is $\frac{\gamma_i[(2-\gamma_i L)(2-\gamma_i\mu)-1]}{2-\gamma_i L-\gamma_i\mu} \geq 0$ for $i = 0, \ldots, N-1$, which is satisfied by any choice of $\gamma_i L \in (0, \overline{\gamma L}_1(\kappa)]$ (see Lemma 4.3). By taking the minimum gradient norm and reordering:

$$
\begin{aligned}
f_0 - f_N &\geq \min_{0\leq i\leq N}\left\{\frac{\|g_i\|^2}{2L}\right\}\Big[\sum_{i=0}^{N-1} L\Big(\frac{\gamma_i L}{L-\mu} + \frac{\gamma_i L\,\gamma_i\mu - 2\gamma_i\mu + \gamma_i L}{L-\mu}\Big)\delta(\gamma_i \leq \tfrac{1}{L}) + \\
&\quad + \Big(\frac{\gamma_i L}{2-\gamma_i L-\gamma_i\mu} + \frac{\gamma_i L[(2-\gamma_i L)(2-\gamma_i\mu)-1]}{2-\gamma_i L-\gamma_i\mu}\Big)\delta(\gamma_i > \tfrac{1}{L})\Big] \\
&= \min_{0\leq i\leq N}\left\{\frac{\|g_i\|^2}{2L}\right\}\sum_{i=0}^{N-1}\gamma_i L\, p(\gamma_i L, \gamma_i\mu),
\end{aligned}
$$

with $p(\gamma_i L, \gamma_i\mu)$ defined in (2.7). Furthermore, the following particular rates hold:
1. Case $\gamma_i L \in (0, 1]$:

$$
\tfrac{1}{2L}\min_{0\leq i\leq N}\left\{\|g_i\|^2\right\} \leq \min\left\{\frac{f_0 - f_N}{\sum_{i=0}^{N-1}\gamma_i L\big(2+\frac{\gamma_i L\gamma_i\mu}{\gamma_i L-\gamma_i\mu}\big)}, \frac{f_0 - f_*}{1+\sum_{i=0}^{N-1}\gamma_i L\big(2+\frac{\gamma_i L\gamma_i\mu}{\gamma_i L-\gamma_i\mu}\big)}\right\};
$$

2. Case $\gamma_i L \in [1, \overline{\gamma L}_1(\kappa)]$:

$$
\tfrac{1}{2L}\min_{0\leq i\leq N}\left\{\|g_i\|^2\right\} \leq \min\left\{\frac{f_0 - f_N}{\sum_{i=0}^{N-1}\frac{\gamma_i L(2-\gamma_i L)(2-\gamma_i\mu)}{2-\gamma_i L-\gamma_i\mu}}, \frac{f_0 - f_*}{1+\sum_{i=0}^{N-1}\frac{\gamma_i L(2-\gamma_i L)(2-\gamma_i\mu)}{2-\gamma_i L-\gamma_i\mu}}\right\}. \square
$$

Proof (**Theorem 2.4**) The interval $(0, \overline{\gamma L}_1(\kappa)]$ is addressed by the proof of Theorem 2.5, so it remains to prove the rates for constant stepsizes in the range

$$
[\overline{\gamma L}_1(\kappa), 2) = \bigcup_{k=1}^{N-1}[\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa)) \bigcup [\overline{\gamma L}_N(\kappa), 2).
$$

1. **Case** $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa))$**,** with some $k \in \{1, \ldots, N-1\}$. To inequality (GN4SD) we append $(N-k-1)$ inequalities (N4SD), written for indices $i = k+1, \ldots, N-1$. Then we get (using the usual notation with γ, L, μ):

$$
\begin{aligned}
\frac{f_0 - f_N}{\gamma} &\geq \frac{-(1-\gamma L)^2(1-\gamma\mu)^2 T_{k+1}(\gamma L, \kappa)}{(1-\gamma\mu)^2-(1-\gamma L)^2}\frac{\|g_k\|^2}{2} + \frac{T_k(\gamma L, \kappa)+\gamma L-\gamma\mu}{(1-\gamma\mu)^2-(1-\gamma L)^2}\frac{\|g_{k+1}\|^2}{2} + \\
&\frac{(2-\gamma L)(2-\gamma\mu)-1}{2-\gamma L-\gamma\mu}\frac{\|g_{k+1}\|^2}{2} + \frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}\sum_{i=k+2}^{N-1}\frac{\|g_i\|^2}{2} + \frac{1}{2-\gamma L-\gamma\mu}\frac{\|g_N\|^2}{2},
\end{aligned}
$$

simplifying to

$$\frac{f_0-f_N}{\gamma}\geq\frac{-(1-\gamma L)^2(1-\gamma\mu)^2T_{k+1}(\gamma L,\kappa)}{(1-\gamma\mu)^2-(1-\gamma L)^2}\frac{\|g_k\|^2}{2}+\frac{T_k(\gamma L,\kappa)}{(1-\gamma\mu)^2-(1-\gamma L)^2}\frac{\|g_{k+1}\|^2}{2}+\\ \frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}\frac{\|g_{k+1}\|^2}{2}+\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}\sum_{i=k+2}^{N-1}\frac{\|g_i\|^2}{2}+\frac{1}{2-\gamma L-\gamma\mu}\frac{\|g_N\|^2}{2},$$

which is of type (♣) since all gradient norms have positive coefficients because $T_k(\gamma L,\kappa)\geq 0$ and $T_{k+1}(\gamma L,\kappa)<0$. By taking the minimum gradient norm:

$$\min_{k\leq i\leq N}\left\{\frac{\|g_i\|^2}{2L}\right\}\leq\frac{f_0-f_N}{\gamma L\,P_N(\gamma L,\gamma\mu)},$$

where

$$P_N(\gamma L,\gamma\mu)=\frac{-(1-\gamma L)^2(1-\gamma\mu)^2T_{k+1}(\gamma L,\kappa)+T_k(\gamma L,\kappa)+\gamma L-\gamma\mu}{(1-\gamma\mu)^2-(1-\gamma L)^2}+\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}(N-k-1).$$

Using algebraic manipulations one can check the following equivalent expressions for $P_N(\gamma L,\gamma\mu)$ discussed in Section 2.3:

$$P_N(\gamma L,\gamma\mu)=\frac{\frac{-1+(1-\gamma L)^{-2k}}{\gamma L[1-(1-\gamma L)^2]}-\frac{-1+(1-\gamma\mu)^{-2k}}{\gamma\mu[1-(1-\gamma\mu)^2]}}{\frac{1}{1-(1-\gamma L)^2}-\frac{1}{1-(1-\gamma\mu)^2}}+\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}(N-k);$$

$$P_N(\gamma L,\gamma\mu)=\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}\left[N-\frac{-\gamma\mu\gamma L}{\gamma L-\gamma\mu}\left[\frac{\frac{-1+(1-\gamma\mu)^{-2k}}{1-(1-\gamma\mu)^2}-k}{\gamma\mu}-\frac{\frac{-1+(1-\gamma L)^{-2k}}{1-(1-\gamma L)^2}-k}{\gamma L}\right]\right].$$

The expression from (2.6) is obtained from the identity:

$$\frac{\frac{-1+(1-\gamma\mu)^{-2k}}{1-(1-\gamma\mu)^2}-k}{\gamma\mu}-\frac{\frac{-1+(1-\gamma L)^{-2k}}{1-(1-\gamma L)^2}-k}{\gamma L}=\sum_{i=0}^{k}\left[\frac{-1+(1-\gamma\mu)^{-2i}}{\gamma\mu}-\frac{-1+(1-\gamma\mu)^{-2i}}{\gamma\mu}\right],$$

where in the r.h.s. are summed up quantities $T_i(\gamma L,\kappa)$ with $i=\{0,1,\ldots,k\}$, i.e., for all indices for which they are positive. Therefore, the r.h.s. can be rewritten using the positive quantities over N terms such as in formula (2.6).

2. **Case** $\gamma L\in[\overline{\gamma L}_N(\kappa),2)$**.** From Lemma 4.12 with $k=N-1$ it holds:

$$\frac{f_0-f_N}{\gamma}\geq E_N(1-\gamma L)\frac{\|g_N\|^2}{2}+\frac{(1-\gamma\mu)^2\,T_N(\gamma L,\kappa)}{(\gamma L-\gamma\mu)^2}\frac{\|g_N-(1-\gamma L)g_{N-1}\|^2}{2},$$

where the mixed term can be neglected since $T_N(\gamma L,\kappa)\geq 0$ and obtain:

$$f_0-f_N\geq\left[-1+(1-\gamma L)^{-2N}\right]\frac{\|g_N\|^2}{2L},$$

which is of type (♣) and implies the linear regime from Theorem 2.4. □

4.5 Proofs of performance bounds for strongly convex functions

If $\mu > 0$, then $\overline{\gamma L}_\infty(\kappa) = \frac{2}{1+\kappa} < 2$.

Proof (**Theorem 2.3**) We split the proof over stepsize intervals:

$$(0,2) = \big(0, \overline{\gamma L}_1(\kappa)\big] \bigcup \bigcup_{k=1}^{N-1} \big[\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa)\big) \bigcup \big[\overline{\gamma L}_N(\kappa), \tfrac{2}{1+\kappa}\big) \bigcup \big[\tfrac{2}{1+\kappa}, 2\big).$$

1. **Case** $\gamma L \in \big(0, \overline{\gamma L}_1(\kappa)\big]$. By telescoping (4.4) for indices $i = 0, \dots, N-1$:

$$\frac{f_0 - f_N}{\gamma} \geq E_N(\eta) \frac{\|g_N\|^2}{2} + \sum_{i=0}^{N-1} \frac{-1 + (\eta+\rho)E_{i+1}(\eta)}{\eta - \rho} \frac{\|g_{i+1} - \eta g_i\|^2}{2},$$

where all mixed terms have nonnegative coefficients because $\gamma L \in \big(0, \overline{\gamma L}_1(\kappa)\big]$ (see Lemma 4.8). Hence they can be neglected to obtain the lower bound of type (♣):

$$\frac{f_0 - f_N}{\gamma} \geq E_N(\eta) \frac{\|g_N\|^2}{2}. \tag{4.10}$$

2. **Case** $\gamma L \in \big[\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa)\big)$, with some $k \in \{1, \dots, N-1\}$. By telescoping (4.4) for indices $i = k+1, \dots, N-1$:

$$\begin{aligned}\frac{f_{k+1} - f_N}{\gamma} \geq & -E_{k+1}(\eta) \frac{\|g_{k+1}\|^2}{2} + E_N(\eta) \frac{\|g_N\|^2}{2} \\ & + \sum_{i=k+1}^{N-1} \frac{-1+(\eta+\rho)E_{i+1}(\eta)}{\eta-\rho} \frac{\|g_{i+1} - \eta g_i\|^2}{2},\end{aligned}$$

with nonnegative weights of the mixed terms since $\gamma L \in \big(0, \overline{\gamma L}_{k+1}(\kappa)\big)$ (see Lemma 4.8). After appending it to (4.7) from Lemma 4.13, we obtain:

$$\begin{aligned}\frac{f_0 - f_N}{\gamma} \geq E_N(\eta) \frac{\|g_N\|^2}{2} + \frac{-\rho^2 T_{k+1}(\rho,\eta)}{(\eta-\rho)^2} \frac{\|g_{k+1} - \eta g_k\|^2}{2} + \\ \sum_{i=k+1}^{N-1} \frac{-1 + (\eta+\rho)E_{i+1}(\eta)}{\eta - \rho} \frac{\|g_{i+1} - \eta g_i\|^2}{2}.\end{aligned}$$

Since $T_{k+1}(\rho, \eta) < 0$, all mixed terms can be neglected to get a (♣)-type inequality:

$$\frac{f_0 - f_N}{\gamma} \geq E_N(\eta) \frac{\|g_N\|^2}{2}. \tag{4.11}$$

3. **Case** $\gamma L \in \left[\overline{\gamma L}_N(\kappa), \frac{2}{1+\kappa}\right)$**.** From Lemma 4.12 with $k = N-1$ it holds:

$$\frac{f_0 - f_N}{\gamma} \geq E_N(\rho)\frac{\|g_N\|^2}{2} + \frac{\eta^2\, T_N(\rho,\eta)}{(\eta-\rho)^2}\frac{\|g_N - \rho g_{N-1}\|^2}{2}.$$

Since $T_N(\gamma L, \kappa) \geq 0$, by neglecting the mixed term we obtain a (♣)-type inequality:

$$\frac{f_0 - f_N}{\gamma} \geq E_N(\rho)\frac{\|g_N\|^2}{2}. \tag{4.12}$$

4. **Case** $\gamma L \in \left[\frac{2}{1+\kappa}, 2\right)$**.** By telescoping Lemma 4.7 for $i = 0, 1, \ldots, N-1$ it holds:

$$\frac{f_0 - f_N}{\gamma} \geq E_N(\rho)\frac{\|g_i\|^2}{2} + \sum_{i=0}^{N-1}\frac{1-(\eta+\rho)E_{i+1}(\rho)}{\eta-\rho}\frac{\|g_{i+1} - \rho g_i\|^2}{2},$$

where the mixed terms can be neglected and then the inequality reduces to (4.12). Since the transition between the two linear regimes is given by the sign of $T_N(\rho,\eta) = E_N(\eta) - E_N(\rho)$, we can glue together expressions (4.10), (4.11) and (4.12):

$$\frac{f_0 - f_N}{\gamma} \geq \min\left\{E_N(\eta)\,,\; E_N(\rho)\right\}\frac{\|g_N\|^2}{2},$$

equivalent to:

$$f_0 - f_N \geq \min\left\{\frac{-1+(1-\gamma\mu)^{-2N}}{\mu}\,,\; \frac{-1+(1-\gamma L)^{-2N}}{L}\right\}\frac{\|g_N\|^2}{2},$$

which implies the bound (2.3) and the complementary one:

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 \leq \frac{f(x_0) - f(x_N)}{\min\left\{\frac{L}{\mu}\left[-1+(1-\gamma\mu)^{-2N}\right]\,,\; -1+(1-\gamma L)^{-2N}\right\}}.$$

□

4.6 Results for convex functions

Corollary 4.14 *(Multistep descent for convex functions) Let $f \in \mathcal{F}_{0,L}$. Consider $k+1 \geq 1$ iterations of (GD) with stepsize $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa))$, satisfying $\gamma L > 1$, starting from x_0. Then the following inequality holds, with positive coefficients of gradient norms:*

$$f(x_0) - f(x_{k+1}) \geq \left[2(k+1)\frac{-(1-\gamma L)^2}{2-\gamma L} + \sum_{i=0}^{k}(1-\gamma L)^{-2i}\right]\frac{\|\nabla f(x_k)\|^2}{2L} +$$

$$+\left[2(k+1)\frac{1}{2-\gamma L}-\sum_{i=0}^{k}(1-\gamma L)^{-2i}\right]\frac{\|\nabla f(x_{k+1})\|^2}{2L}. \tag{G4SD}$$

Moreover, (G4SD) *is valid on the extended stepsize range* $\gamma L \in (1,2)$.

Proof It results by taking $\eta = 1$ in (GN4SD). □

(G4SD) is a multistep generalization of the descent lemma (4SD) from [34].

Proof (**Theorem 2.2**) Letting $\mu \nearrow 0$, all terms summed in (2.6) become zero, simplifying to $P_N(\gamma L, 0) = 2N$, so that all sublinear regimes corresponding to stepsizes up to $\overline{\gamma L}_N(0)$ share the same expression. For stepsizes $\gamma L \geq \overline{\gamma L}_N(0)$, it holds the linear regime following the expression in the denominator $E_N(1-\gamma L) = \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}$. Additionally, the thresholds are determined by the roots of:

$$T_k(\gamma L, 0) = 2k - \frac{-1+(1-\gamma L)^{-2k}}{\gamma L},\ \forall k = 1,\dots,N,$$

hence the rate can be expressed as a minimum in the denominator of (2.1). Same bound is recovered to the limit $\mu \searrow 0$ in strongly convex functions' rate (2.3). □

5 Tightness of performance bounds

The tightness of convergence rates is typically assessed by identifying a worst-case function matching the upper bound when iterating the algorithm. This function can be determined either analytically or by providing an interpolable set of triplets. Table 1 summarizes the various demonstrations of tightness based on different stepsizes ranges.

Remark 5.1 The exactness of all proved convergence rates can be confirmed numerically by solving the associated performance estimation problems (PEPs), for example by using the specialized software packages PESTO [31] (for Matlab) or PEPit [18] (for Python). In this section we propose analytical proofs.

For stepsizes in the range $\gamma L \in [\overline{\gamma L}_N(\kappa), 2)$, a single performance bound applies across all function classes (weakly convex, convex and strongly convex). As shown in Proposition 5.2, this unified bound is proven to be tight using the same worst-case function $L\frac{\|x\|^2}{2}$, which, for simplicity, can be reduced to one dimension.

Proposition 5.2 *(Tightness for* $\gamma L \in [\overline{\gamma L}_N(\kappa), 2)$*) Let* $L > 0$*, N be a positive integer,* $\gamma L \in [\overline{\gamma L}_N(\kappa), 2)$ *and* $f_L(x) := L\frac{\|x\|^2}{2}$*. Then after applying N iterations of* (GD) *on function* $f = f_L$ *with constant stepsize* γL*, starting from* x_0*, it holds:*

$$\tfrac{1}{2L}\min_{0\leq i\leq N}\{\|\nabla f(x_i)\|^2\} = \tfrac{1}{2L}\|\nabla f(x_N)\|^2 = \frac{f(x_0)-f_*}{(1-\gamma L)^{-2N}}.$$

Proof One can directly check that f_L satisfies the necessary conditions for equality in Lemma 4.12: $g_N = (1-\gamma L)g_{N-1}$ and $f_0 - f_N = [-1+(1-\gamma L)^{-2N}]\frac{\|g_N\|^2}{2L}$. □

For the remaining of the stepsize interval, $\gamma L \in (0, \overline{\gamma L}_N(\kappa))$, the proofs are organized as follows: Section 5.1 covers constant stepsizes for convex and strongly convex functions; Section 5.3 addresses variable stepsizes below $\overline{\gamma L}_1(\kappa)$ for convex and weakly convex functions; Section 5.4 focuses on the specific range of constant stepsizes $(1, \overline{\gamma L}_N(\kappa))$ for weakly convex functions.

5.1 Tightness for convex and strongly convex functions

Proposition 5.3 *(Tightness of Theorem 2.3) Let $L > 0$, $\mu \in (0, L)$, $\gamma L \in (0, 2)$ and N be a positive integer. Then there exists x_0 and $f \in \mathcal{F}_{\mu,L}$ such that after applying N iterations of (GD) on function f with constant stepsize $\gamma L \in (0, 2)$, starting from x_0, it holds:*

$$\tfrac{1}{2L}\|\nabla f(x_N)\|^2 = \frac{f(x_0) - f_*}{1 + \gamma L \min\left\{\frac{-1+(1-\gamma\mu)^{-2N}}{\gamma\mu}\,,\ \frac{-1+(1-\gamma L)^{-2N}}{\gamma L}\right\}}.$$

Proof We define the following worst-case function depending on the stepsizes γL: if $\gamma L \in [\overline{\gamma L}_N(\kappa), 2)$, then it is given by the function f_L (defined in Proposition 5.2); if $\gamma L \in (0, \overline{\gamma L}_N(\kappa))$, then it is the function φ given below. Let $\Delta > 0$ and define the parameter $\tau := \sqrt{\frac{2}{L}\frac{\Delta}{1+\frac{L}{\mu}[-1+(1-\gamma\mu)^{-2N}]}}$. The function $\varphi\colon \mathbb{R} \to \mathbb{R}$ is given by:

$$\varphi(x) := \begin{cases} \mu\frac{x^2}{2} + (L-\mu)\tau|x| - (L-\mu)\frac{\tau^2}{2}, & \text{if } |x| \geq \tau; \\ L\frac{x^2}{2}, & \text{if } |x| < \tau. \end{cases}$$

Note that $\varphi \in \mathcal{F}_{\mu,L}$, $x_* = \arg\min \varphi(x) = 0$, $\varphi(x_*) = 0$ and

$$\nabla\varphi(x) := \begin{cases} \mu x + (L-\mu)\tau\,\mathrm{sgn}(x), & \text{if } |x| \geq \tau; \\ Lx, & \text{if } |x| < \tau. \end{cases}$$

One can check that initializing the iterates at $x_0 := \tau[1+\frac{L}{\mu}\left(-1+(1-\gamma\mu)^{-N}\right)]$, with $\varphi(x_0) = \Delta$, and running N iterations of (GD), the following holds: all iterations belong to the branch $x \geq \tau$, the final iterate is exactly $x_N = \tau$, having the gradient $\nabla\varphi(x_N) = L\tau$ and thus reaching the bound claimed in the first term. □

This one-dimensional worst-case example is inspired from [32, Section 4.1.2].

Proposition 5.4 *(Tightness of Theorem 2.2) Let $L > 0$, $\gamma L \in (0, 2)$ and N be a positive integer. Then there exists x_0 and $f \in \mathcal{F}_{0,L}$ such that after applying N iterations of (GD) on function f with constant stepsize $\gamma L \in (0, 2)$, starting from x_0, it holds:*

$$\tfrac{1}{2L}\|\nabla f(x_N)\|^2 = \frac{f(x_0) - f_*}{\min\left\{1 + 2N\gamma L\,,\ (1-\gamma L)^{-2N}\right\}}.$$

Proof With $\Delta > 0$, let $\tau := \sqrt{\frac{2}{L}\frac{\Delta}{1+2N\gamma L}}$ and $x_0 = \tau(1+N\gamma L)$. We define $\varphi\colon \mathbb{R} \to \mathbb{R}$, such that $\varphi(x_0) = \Delta$, $\varphi(x_*) = 0$ and

$$\varphi(x) := \begin{cases} L\tau|x| - L\frac{\tau^2}{2}, & \text{if } |x| \geq \tau; \\ L\frac{x^2}{2}, & \text{if } |x| < \tau, \end{cases} \quad \nabla\varphi(x) := \begin{cases} L\tau\,\mathrm{sgn}(x), & \text{if } |x| \geq \tau; \\ Lx, & \text{if } |x| < \tau. \end{cases}$$

Then we select $f = \varphi$ or $f = f_L$ (see Proposition 5.2) whether $\gamma L < \overline{\gamma L}_N(0)$ or $\gamma L \geq \overline{\gamma L}_N(0)$, respectively. □

5.2 Removing the optimal point in tightness analysis for smooth weakly convex functions

Lemma 5.5 shows that by incorporating inequalities (3.3) for all iterates $i \in \mathcal{I}$ the upper bounds remain tight and provides an existence guarantee of an interpolating function with global minimum f_*. It generalizes Theorem 7 from [15], obtained for the particular choices $\mu = -L$ and $\mu = 0$, to any $\mu \leq 0$.

Lemma 5.5 *Let $\mathcal{T} = \{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ be $\mathcal{F}_{\mu,L}$-interpolable, with $\mu \in (-\infty, 0]$. There exists at least one interpolating function with a finite global minimum*

$$f_* = \min_{x\in\mathbb{R}^d} f(x) = \min_{i\in\mathcal{I}} \left\{ f_i - \tfrac{1}{2L}\|g_i\|^2 \right\}. \tag{5.1}$$

A global minimizer is given by $x_* = x_{i_*} - \frac{1}{L}g_{i_*}$, where $i_* \in \arg\min_{i\in\mathcal{I}} \left\{ f_i - \frac{1}{2L}\|g_i\|^2 \right\}$.

Proof The proof consists on algebraic manipulations, see Appendix B.2. □

Lemma 5.6 *Assume an $\mathcal{F}_{\mu,L}$-interpolable set $\mathcal{T} = \{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ satisfying the condition*

$$N \in \arg\min_{i\in\mathcal{I}} \left\{ f_i - \tfrac{1}{2L}\|g_i\|^2 \right\}. \tag{5.2}$$

Then proving tightness of the bound

$$\tfrac{1}{2L} \min_{0\leq i\leq N} \{\|\nabla f(x_i)\|^2\} \leq \frac{f(x_0) - f_*}{1 + P_N(\gamma L, \gamma\mu)}$$

reduces to showing tightness of the bound

$$\tfrac{1}{2L} \min_{0\leq i\leq N} \{\|\nabla f(x_i)\|^2\} \leq \frac{f(x_0) - f(x_N)}{P_N(\gamma L, \gamma\mu)}. \tag{5.3}$$

Proof Condition (5.2) implies that (5.1) from Lemma 5.5 holds with $N = i_*$, hence there exists a function for which $f(x_N) - f_* \geq \frac{1}{2L}\|\nabla f(x_N)\|^2$. This inequality produces the shift in the denominator when appended to rate (5.3). □

Further on, in all tightness proofs we show that condition (5.2) holds and therefore we reduce the analysis to only prove the equality in (5.3).

5.3 Tightness analysis for weakly convex functions and stepsizes below $\overline{\gamma L}_1$

Range $\gamma_i L \in (0, 1]$. We show a one dimensional worst-case function example which extends to weakly convex functions the one from [1, Proposition 4].

Proposition 5.7 *(Tightness for $\gamma L \in (0, 1]$) Let $L > 0$, $\mu \in (-\infty, 0)$, N be a positive integer and $\gamma_i L \in (0, 1]$, with $i = 0, \dots, N-1$. Then there exists x_0 and $f \in \mathcal{F}_{\mu,L}$ such that after applying N iterations of (GD) on function f with stepsizes $\gamma_i L$, starting from x_0, it holds:*

$$\frac{1}{2L} \min_{0 \le i \le N} \{\|\nabla f(x_i)\|^2\} = \frac{f(x_0) - f_*}{1 + \sum_{i=0}^{N-1} \gamma_i L \left(2 + \frac{\gamma_i L \gamma_i \mu}{\gamma_i L - \gamma_i \mu}\right)}.$$

Proof Motivated by the necessary conditions $g_i = g_{i+1}$ in Lemma 4.2, a one dimensional function with constant gradients U at the iterates is built, where U is the square root of the minimum gradient norm: $U^2 := \frac{\Delta}{\frac{1}{2L}\sum_{i=0}^{N-1} \gamma_i L (2 + \frac{\gamma_i L \gamma_i \mu}{\gamma_i L - \gamma_i \mu})}$, with $\Delta > 0$. The equality case in Lemma 4.2 implies (together with $g_i = U$, $\forall i = 0, 1, \dots, N$):

$$\frac{f_i - f_{i+1}}{\Delta} = \frac{\gamma_i L \left(2 + \frac{\gamma_i L \gamma_i \mu}{\gamma_i L - \gamma_i \mu}\right)}{\sum_{j=0}^{N-1} \gamma_j L \left(2 + \frac{\gamma_j L \gamma_j \mu}{\gamma_j L - \gamma_j \mu}\right)} \Leftrightarrow \frac{f_i}{\Delta} = \frac{\sum_{j=i}^{N-1} \gamma_j L \left(2 + \frac{\gamma_j L \gamma_j \mu}{\gamma_j L - \gamma_j \mu}\right)}{\sum_{j=0}^{N-1} \gamma_j L \left(2 + \frac{\gamma_j L \gamma_j \mu}{\gamma_j L - \gamma_j \mu}\right)}.$$

Without loss of generality, we set $f_N = 0$, $f_0 = \Delta$ and $x_N = 0$ and define the triplets $\{(x_i, g_i, f_i)\}_{i \in \{0,\dots,N\}}$, with $g_i = U$ and $x_i = U \sum_{j=i}^{N-1} \gamma_j$. Consider the following points lying between consecutive iterates:

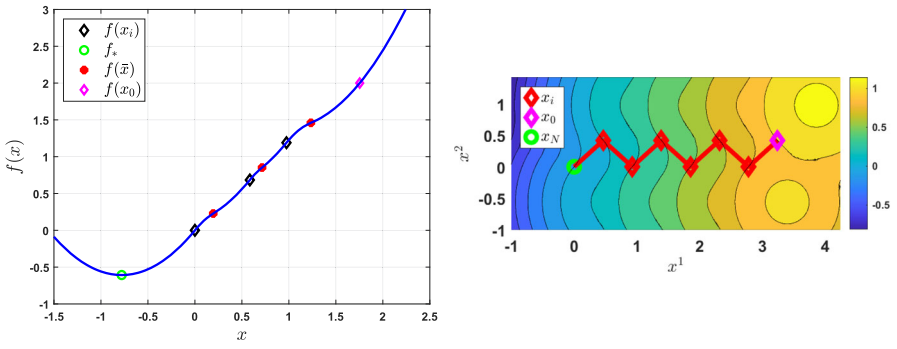
$$\bar{x}_i := x_i - \tfrac{-\mu}{L-\mu} \gamma_i U \in [x_{i+1}, x_i], \quad i = 0, 1, \dots, N-1.$$

A worst-case function interpolating the triplets is the piecewise quadratic $f : \mathbb{R} \to \mathbb{R}$, with alternating curvature between μ and L with inflection points x_i and $\bar{x}_i$, constant gradients at the iterations and decreasing function values, where $i \in \{0, \dots, N-1\}$:

$$f(x) = \begin{cases} \frac{L}{2}(x - x_N)^2 + U(x - x_N) + f_N, & \text{if } x \in (-\infty, x_N]; \\ \frac{\mu}{2}(x - x_{i+1})^2 + U(x - x_{i+1}) + f_{i+1}, & \text{if } x \in [x_{i+1}, \bar{x}_i]; \\ \frac{L}{2}(x - x_i)^2 + U(x - x_i) + f_i, & \text{if } x \in [\bar{x}_i, x_i]; \\ \frac{L}{2}(x - x_0)^2 + U(x - x_0) + f_0, & \text{if } x \in [x_0, \infty). \end{cases} \tag{5.4}$$

□

An illustration of the one-dimensional worst-case construction from Proposition 5.7 is given in Figure 5a.



(a) (b)

Fig. 5 Worst-case examples for Theorems 2.4 and 2.5 with $\gamma L \leq \overline{\gamma L}_1(\kappa)$. (**a**) One dimensional worst-case function example corresponding to short stepsizes $\gamma_i L \in (0, 1]$ (piecewise quadratic; see Proposition 5.7). Setup: $N = 3$, $f_0 - f_* = 2$, $L = 2$, $\mu = -4$, $\gamma_0 L = 1$, $\gamma_1 L = 0.5$, $\gamma_2 L = 0.75$. (**b**) Function example for constrant stepsize $\gamma L \in (1, \overline{\gamma L}_1(\kappa)]$. Setup: $L = 1$, $\mu = -0.5$, $\gamma L = 1.5$, $N = 7$. Component x^1 is successively reduced, while x^2 alternates between two values, due to alternating gradients g_i and g_{i+1}

Corollary 5.8 *(Tightness of Corollary 2.8 for convex functions) Let $L > 0$, N be a positive integer and $\gamma_i L \in (0, \frac{3}{2}]$, with $i = 0, \ldots, N-1$. Then there exists x_0 and $f \in \mathcal{F}_{0,L}$ such that after applying N iterations of (GD) on function f with stepsizes $\gamma_i L$, starting from x_0, it holds:*

$$\frac{1}{2L}\|\nabla f(x_N)\|^2 = \frac{f(x_0) - f_*}{1 + 2\sum_{i=0}^{N-1} \gamma_i L}.$$

Proof The proof results by setting $\mu = 0$ in Proposition 5.7 to obtain a linear function between iterations, extended outside by a quadratic of curvature L. □

Range $\gamma_i L \in (1, \overline{\gamma L}_1(\kappa)]$**.** We build a set of triplets matching the corresponding bounds from Theorems 2.4 and 2.5 and interpolating an $\mathcal{F}_{\mu,L}$-function, implying the validity of all interpolation inequalities relating the $N+1$ iterations. The upper bounds in Theorem 2.5 are obtained by analyzing only consecutive iterations, from which we track equality conditions collected in Proposition 5.9.

Proposition 5.9 *Let $L > 0$, $\mu \in (-\infty, 0)$, N be a positive integer and $\gamma_i L \in (1, \overline{\gamma L}_1(\kappa))$, with $i = 0, \ldots, N-1$. Then any worst-case function $f \in \mathcal{F}_{\mu,L}$ matching the bound from Theorem 2.5 satisfies the following identities on the triplets $\mathcal{T} = \{(x_i, g_i, f_i)\}_{i \in \{0,\ldots,N\}}$, with $x_{i+1} = x_i - \gamma_i g_i$:*

$$\|g_i\|^2 = U^2 := \frac{f_0 - f_N}{\frac{1}{2L}\sum_{j=0}^{N-1} \frac{\gamma_j L(2-\gamma_j L)(2-\gamma_j \mu)}{2-\gamma_j L-\gamma_j \mu}}, \quad \forall i = 0, \ldots, N; \tag{5.5}$$

$$c(\gamma_i L, \gamma_i \mu) := \frac{\langle g_i, g_{i+1}\rangle}{U^2} = \frac{1 + (1-\gamma_i \mu)(1-\gamma_i L)}{2 - \gamma_i L - \gamma_i \mu}, \quad \forall i = 0, \ldots, N-1; \tag{5.6}$$

$$f_i = f_N + \frac{U^2}{2L}\sum_{j=i}^{N-1}\frac{\gamma_j L(2-\gamma_j\mu)(2-\gamma_j L)}{2-\gamma_j L-\gamma_j\mu}, \quad \forall i = 0,\dots,N. \tag{5.7}$$

In particular, a two-dimensional function $f:\mathbb{R}^2\to\mathbb{R}$ satisfies these necessary conditions if, for any suitable g_0, the following holds:

$$g_{i+1} = R\big((-1)^{i+1}\theta(\gamma_i L,\gamma_i\mu)\big)g_i, \quad \forall i = 0,1,\dots,N-1, \tag{5.8}$$

where R is the two-dimensional rotation matrix and

$$\theta(\gamma_i L,\gamma_i\mu) := \arccos\big(c(\gamma_i L,\gamma_i\mu)\big). \tag{5.9}$$

Moreover,

$$N \in \operatorname*{arg\,min}_{0\le i\le N}\left\{f_i - \tfrac{1}{2L}\|g_i\|^2\right\}. \tag{5.10}$$

Proof See Appendix C.1. □

Proposition 5.10 *(Tightness for constant stepsize $\gamma L\in(1,\overline{\gamma L}_1)$) Let $L>0$, $\mu\in(-\infty,0)$, N be a positive integer and $\gamma L\in(1,\overline{\gamma L}_1(\kappa))$. Then there exists x_0 and $f\in\mathcal{F}_{\mu,L}$ such that after applying N iterations of* (GD) *on function f with constant stepsize γL, starting from x_0, it holds:*

$$\tfrac{1}{2L}\min_{0\le i\le N}\{\|\nabla f(x_i)\|^2\} = \frac{f(x_0)-f_*}{1+\gamma L\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}N}.$$

Proof We construct a two-dimensional worst-case function f: $\mathbb{R}^2\to\mathbb{R}$. From Proposition 5.9, the rotation angle between the gradients $\theta(\gamma L,\gamma\mu)$ is constant, hence they alternate between odd and even iterations such that:

$$\|g_i\|^2 = \langle g_i, g_{i+(2\mathbb{N})}\rangle = U^2 \text{ and } \langle g_i, g_{i+(2\mathbb{N}+1)}\rangle = cU^2 \qquad \forall i = 0,\dots,N-1.$$

Consider the set of triplets $\mathcal{T}_{\mathcal{I}} := \{(x_i, g_i, f_i)\}_{i\in\{0,\dots,N\}}$ defined as follows: (i) since g_0 can be set (almost) arbitrary, we take $g_i = U\left[\sqrt{\frac{1+c}{2}};\ (-1)^i\sqrt{\frac{1-c}{2}}\right]^T$; (ii) by setting $x_N=[0,0]^T$, the iterations are given by $x_i=\gamma\sum_{j=i}^{N-1}g_j$; (iii) by setting $f_N=0$ the function values simplify to $f_i = f_0(1-\frac{i}{N})$. In Appendix C.1 we demonstrate that $\mathcal{T}_{\mathcal{I}}$ is $\mathcal{F}_{\mu,L}$-interpolable. □

Figure 5b shows a numerical example of the triplets from Proposition 5.10's proof. Excepting initial point x_0, all subsequent iterations are inflection points where the curvature transitions from μ to L. This pattern is reversed between odd and even iterations and the decrease mainly occurs to component x^1.

Remark 5.11 The numerical example from Figure 5b is obtained from the lower interpolation function derived in [29, Theorem 3.14], originally defined for smooth strongly convex functions. This interpolation function can be extended to arbitrary lower curvatures μ using the minimal curvature subtraction argument employed in the proof of Theorem 3.2 (refer to Remark 3.13 in [29]). Given triplets $\mathcal{T}_{\mathcal{I}} := \{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$, this function is the solution of the quadratically constrained optimization problem (QCQP):

$$f(x) := \min_{g\in\mathbb{R}^2} \max_{0\leq i\leq N} \left\{ f_i + \langle g_i\,,\, x-x_i\rangle + \tfrac{\mu}{2}\|x-x_i\|^2 + \tfrac{1}{2(L-\mu)}\|g-g_i-\mu(x-x_i)\|^2 \right\}.$$

When using non-constant stepsizes $\gamma_i L \in (1, \overline{\gamma L}_1(\kappa)]$, based on the triplets construction from Proposition 5.9 we state in Conjecture 1 the tightness of the convergence rate from Theorem 2.5. The conjecture relies on numerically verification that all interpolation conditions hold (see Footnote 1).

Conjecture 1 *(Lower bound for $\gamma_i L \in (1, \overline{\gamma L}_1]$)) There exists a two-dimensional function $f \in \mathcal{F}_{\mu,L}$ interpolating the triplets defined in Proposition 5.9, satisfying*

$$\tfrac{1}{2L} \min_{0\leq i\leq N} \{\|\nabla f(x_i)\|^2\} = \frac{f(x_0) - f(x_N)}{\sum_{i=0}^{N-1} \frac{\gamma_i L(2-\gamma_i L)(2-\gamma_i\mu)}{2-\gamma_i L-\gamma_i\mu}}.$$

5.4 Tightness analysis for weakly convex functions and stepsizes above $\overline{\gamma L}_1$

We firstly show the tightness for the stepsize interval $[\overline{\gamma L}_{N-1}(\kappa), \overline{\gamma L}_N(\kappa))$ in Proposition 5.12, by constructing a two-dimensional quadratic function. Then Proposition 5.14 demonstrates the exactness of our bounds for any stepsize in the range $(1, \overline{\gamma L}_N(\kappa))$, by providing in its proof a three-dimensional example. We also provide alternative, equivalent expressions of denominator $P_N(\gamma L, \gamma\mu)$ from (2.6).

Using $\bar{N}(\gamma L, \kappa)$ (recall Definition 4.5) emphasizes that the analysis of worst-case functions relies on fixing the stepsizes and varying the number of iterations; $\bar{N}(\gamma L, \kappa)$ plays the role of index k delimiting intervals in the proofs of Theorem 2.4 and Theorem 2.3 from Section 4.4 and Section 4.5, respectively.

Proposition 5.12 *(Tightness for $\gamma L \in [\overline{\gamma L}_{N-1}, \overline{\gamma L}_N)$) Let $L>0$, $\mu \in (-\infty, 0)$, N be a positive integer and $\gamma L \in [\overline{\gamma L}_{N-1}(\kappa), \overline{\gamma L}_N(\kappa))$, such that $\bar{N}(\gamma L, \kappa)=N-1$. Then there exists x_0 and $f \in \mathcal{F}_{\mu,L}$ such that after applying N iterations of (GD) on function f with constant stepsize γL, starting from x_0, the corresponding performance bound from Theorem 2.4 holds exactly:*

$$\tfrac{1}{2L} \min_{0\leq i\leq N} \{\|\nabla f(x_i)\|^2\} = \frac{f(x_0) - f_*}{1 + \gamma L \frac{\frac{(1-\gamma L)^2 E_N(1-\gamma L)}{1-(1-\gamma L)^2} - \frac{(1-\gamma\mu)^2 E_N(1-\gamma\mu)}{1-(1-\gamma\mu)^2}}{\frac{1}{1-(1-\gamma L)^2} - \frac{1}{1-(1-\gamma\mu)^2}}},$$

with $E_k(x)$ given in Definition 2.9.

Proof Let $\Delta > 0$ and $U = \sqrt{\frac{2L\,\Delta}{\gamma L P_N(\gamma L, \gamma\mu)}}$, with $P_N(\gamma L, \gamma\mu)$ defined in (2.6). The following performance bound (without f_*)

$$\frac{1}{2L}\min_{0\le i\le N}\{\|\nabla f(x_i)\|^2\} = \frac{f(x_0)-f(x_N)}{\gamma L}\,\frac{\frac{1}{1-(1-\gamma L)^2}-\frac{1}{1-(1-\gamma\mu)^2}}{\frac{(1-\gamma L)^2\,E_N(1-\gamma L)}{1-(1-\gamma L)^2}-\frac{(1-\gamma\mu)^2\,E_N(1-\gamma\mu)}{1-(1-\gamma\mu)^2}}$$

is achieved by the $\mathcal{F}_{\mu,L}$ quadratic function

$$f(x) = \frac{1}{2}(x-x_N)^T\begin{bmatrix} L & 0\\ 0 & \mu\end{bmatrix}(x-x_N) + g_N^\top(x-x_N),$$

with arbitrarily fixed $x_N = [0,0]^T$, $f_N = 0$, $f(x_0) = \Delta$ and

$$g_N := U\sqrt{\frac{[(1-\gamma\mu)^2-1][1-(1-\gamma L)^2]}{(1-\gamma\mu)^2-(1-\gamma L)^2}}\begin{bmatrix}\frac{1-\gamma L}{\sqrt{1-(1-\gamma L)^2}}\\ \frac{1-\gamma\mu}{\sqrt{(1-\gamma\mu)^2-1}}\end{bmatrix}.$$

□

Remark 5.13 The gradients of the worst-case function from Proposition 5.12, evaluated at iterations x_i, decrease in each component by the contraction factors $(1-\gamma L)$ and $(1-\gamma\mu)$, respectively:

$$\nabla f(x_{i+1}) = \begin{bmatrix}1-\gamma L & 0\\ 0 & 1-\gamma\mu\end{bmatrix}\nabla f(x_i).$$

Proving the tightness of Theorem 2.4 for stepsizes $\gamma L \in \left[\overline{\gamma L}_1(\kappa), \overline{\gamma L}_{N-1}(\kappa)\right)$ presents a significant challenge due to the nonlinear term in (2.6), which connects the quadratic function characterizing the first $\bar{N}(\gamma L,\kappa)+1$ iterations (see Proposition 5.12) and the 2D function with alternating gradients directions from Proposition 5.10. This later function dominates the sublinear rate through the leading coefficient $\gamma L\, p(\gamma L, \gamma\mu)$. The three-dimensional worst-case construction from Proposition 5.14's proof also provides an alternative worst-case construction for this regime with stepsizes $\gamma L \in (1, \overline{\gamma L}_1]$.

Proposition 5.14 *(Tightness for $\gamma L \in (1, \overline{\gamma L}_N)$) Let $L > 0$, $\mu \in (-\infty, 0)$, N be a positive integer and $\gamma L \in (1, \overline{\gamma L}_N(\kappa))$. There exists x_0 and $f \in \mathcal{F}_{\mu,L}$ such that after applying N iterations of* (GD) *on function f with constant stepsize γL, starting from x_0, the performance bound from Theorem 2.4 is attained:*

$$\frac{1}{2L}\min_{0\le i\le N}\{\|\nabla f(x_i)\|^2\} = \frac{f(x_0)-f_*}{1+\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}(N-\bar{N})+\frac{\frac{E_{\bar{N}}(1-\gamma L)}{1-(1-\gamma L)^2}-\frac{E_{\bar{N}}(1-\gamma\mu)}{1-(1-\gamma\mu)^2}}{\frac{1}{1-(1-\gamma L)^2}-\frac{1}{1-(1-\gamma\mu)^2}}},$$

where $E_k(x)$ given in Definition 2.9, together with the complementary rate:

$$\frac{1}{2L}\min_{0\leq i\leq N}\{\|\nabla f(x_i)\|^2\} = \frac{f(x_0)-f(x_N)}{\frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}(N-\bar{N}) + \frac{\frac{E_{\bar{N}}(1-\gamma L)}{1-(1-\gamma L)^2}-\frac{E_{\bar{N}}(1-\gamma\mu)}{1-(1-\gamma\mu)^2}}{\frac{1}{1-(1-\gamma L)^2}-\frac{1}{1-(1-\gamma\mu)^2}}}.$$

Proof Let $\Delta > 0$ and $U = \sqrt{\frac{2L\,\Delta}{\gamma L P_N(\gamma L,\gamma\mu)}}$, with $P_N(\gamma L,\gamma\mu)$ defined in (2.6). The value of U denotes the minimum gradient norm over N iterations, i.e., the rate, where Δ accounts for the initial function value gap $f(x_0)-f(x_N)$. The definition of $\bar{N} = \bar{N}(\gamma L,\kappa)$ implies $\gamma L \in [\overline{\gamma L}_{\bar{N}-1}(\kappa), \overline{\gamma L}_{\bar{N}}(\kappa))$, with $1\leq\bar{N}\leq N-1$. We define the set of triplets $\mathcal{T}_{\mathcal{I}} := \{(x_i, g_i, f_i)\}_{i=\{0,1,\ldots,N\}}$ as follows:

$$g_i = U\sqrt{\frac{[(1-\gamma\mu)^2-1][1-(1-\gamma L)^2]}{(1-\gamma\mu)^2-(1-\gamma L)^2}}\begin{bmatrix}\frac{(1-\gamma L)^{i-\bar{N}}}{\sqrt{1-(1-\gamma L)^2}}\\ \frac{(1-\gamma\mu)^{i-\bar{N}}}{\sqrt{(1-\gamma\mu)^2-1}}\\ 0\end{bmatrix} \quad \forall i=0,1,\ldots,\bar{N}+1;$$

$$g_{i+1} = (1-\gamma L)g_i + \gamma(L-\mu)\cos(\theta_{i+1})\begin{bmatrix}0 & \\ & R(\theta_{i+1})\end{bmatrix}g_i \quad \forall i=\bar{N},\ldots,N;$$

$$q_i = (-1)^{i-(\bar{N}+1)}\sqrt{[(1-\gamma\mu)^2-1]\tfrac{1-(1-\gamma L)^{2(i-(\bar{N}+1))}}{1-(1-\gamma L)^2}} \quad \forall i=\bar{N}+1,\ldots,N;$$

$$\theta_i = \arctan(q_i) \quad \forall i=\bar{N}+1,\ldots,N;$$

$$f_i = \Delta\frac{N-i+\frac{\gamma L\gamma\mu}{\gamma L-\gamma\mu}\sum_{j=1}^{\bar{N}-i}T_j(\gamma L,\kappa)}{N+\frac{\gamma L\gamma\mu}{\gamma L-\gamma\mu}\sum_{j=1}^{\bar{N}}T_j(\gamma L,\kappa)} \quad i=0,1,\ldots,N,$$

$f_N = 0$, $x_N = 0$ and $x_{i+1} = x_i - \gamma g_i$, $i = 0,\ldots,N-1$ and $R(\theta) := \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}$ being the two-dimensional rotation matrix of angle θ. For $\mathcal{T}_{\mathcal{I}}$, we show in Appendix C.2 that: (i) it attains the performance bound from Theorem 2.4 and (ii) it is interpolable by a $\mathcal{F}_{\mu,L}$-function with three variables. □

Remark 5.15 If the stepsize is not the threshold $\overline{\gamma L}_{\bar{N}}(\kappa)$, then the first $\bar{N}+1$ gradients have uniquely defined norm and inner products between them. This follows the equality conditions from Lemma 4.11. Restricting them to be two-dimensional, they are unique up to applying orthonormal transformations. The first $\bar{N}+1$ iterations are exactly characterized by the worst-case function from Proposition 5.12, where a third component is set to zero. For the subsequent iterations, the first component remains linearly decreasing in $(1-\gamma L)$, while the others are obtained by applying a certain rotation which, given the decrease in the first component, preserves: (i) the gradient norm to U^2 and (ii) the inner product between consecutive iterations to cU^2 such that:

$$\text{(i): } \|g_{i+1} - (1-\gamma L)g_i\|^2 = U^2\big[1 + (1-\gamma L)^2 - 2(1-\gamma L)c\big];$$
$$\text{(ii): } \langle g_{i+1} - (1-\gamma L)g_i \,,\, g_i\rangle = U^2\big[c - (1-\gamma L)\big],$$

with $c{=}\frac{1+(1-\gamma L)(1-\gamma\mu)}{2-\gamma L-\gamma\mu}$ from Proposition 5.10. The matrix $R(\theta_{i+1})$ rotates the second and third components of g_i by angle θ_i, whose absolute value monotonically increasing towards $\lim_{i\to\infty}|\theta_i| = \arctan(\sqrt{\frac{-\mu}{L}\frac{2-\gamma\mu}{2-\gamma L}})$. These components are then properly scaled to preserve $\|g_{i+1}\|{=}U$. Moreover, as i approaches ∞, the first component vanishes and the gradients start alternating, resembling the same pattern observed for the two-dimensional worst-case described in Proposition 5.10's proof.

6 Conclusion

We have presented a comprehensive worst-case analysis of gradient descent applied to smooth weakly convex, convex and strongly convex functions across the entire range of constant stepsizes $\gamma L \in (0, 2)$.

For convex and strongly convex functions we establish and prove tight performance bounds over ranges previously only covered by conjectures.

In the weakly convex scenario, determining convergence rates is more challenging outside extreme cases already addressed in the literature: $\mu \in \{-\infty, 0\}$. The state-of-the-art analysis of smooth non-necessarily convex functions is extended from the particular case $\mu = -L$ and stepsizes $\gamma L \in (0, \sqrt{3}]$ to encompass full domains.

The proofs are based on incorporating not only consecutive interpolation inequalities but also those connecting iterations at a distance of two, the key idea being detailed in Remark 4.10. A key advantage of these proofs type is the obtaining of new descent lemma type inequalities, independent of the number of iterations N, as they would be in tight convergence analyses using inequalities involving the last iterate. When deriving closed-form expressions for tight performance analysis with constant stepsizes, it is uncommon and challenging to involve nonconsecutive iterations.

While all rates align with the numerical solutions obtained from solving performance estimation problems (PEPs), we demonstrate analytically their tightness over the complete stepsize range $\gamma L \in (0, 2)$.

This analysis yields insights into the optimal constant stepsizes with respect to the demonstrated worst-case scenarios and we further improve upon them by proposing dynamic stepsize schedules which have the key benefit of being independent of the number of iterations.

Acknowledgements The authors thank the anonymous referees and the handling editor for their detailed and constructive comments, which substantially improved the presentation and strengthened the results compared to the initial version of the paper.

Funding This work was supported by the framework of the Global PhD Partnership KU Leuven - UCLouvain.

Code availability The symbolic script to check the derivations and the numerical examples are available on https://github.com/teo2605/GD_tight_rates.

A Properties of stepsize sequences (proofs)

A.1 Stepsize thresholds

Definition A.1 We use the writing $T_k(\gamma L, \kappa) = \sum_{i=1}^{k} \delta_i(\gamma L, \kappa)$, where

$$\delta_i(\gamma L, \kappa) := \frac{2-\kappa\gamma L}{(1-\kappa\gamma L)^{2i}} - \frac{2-\gamma L}{(1-\gamma L)^{2i}}.$$

Proof (**Proposition 2.10**) We show $\frac{dT_k(\gamma L,\kappa)}{d(\gamma L)} \geq 0$ for $\gamma L \in (1, \overline{\gamma L}_{\infty}(\kappa))$, which resumes to prove the positivity of

$$\frac{d\delta_i(\gamma L, \kappa)}{d(\gamma L)} := \frac{-\kappa\left(1 - 2i\frac{2-\kappa\gamma L}{1-\kappa\gamma L}\right)}{(1-\kappa\gamma L)^{2i}} + \frac{1 + 2i\frac{2-\gamma L}{\gamma L-1}}{(1-\gamma L)^{2i}}.$$

1. **Case** $\kappa \geq 0$**:** we lower bound $\frac{d\delta_i(\gamma L,\kappa)}{d(\gamma L)}$ by neglecting the denominators:

$$\begin{aligned}\delta_i(\gamma L, \kappa) &\geq -\kappa\Big(1 - 2i\frac{2-\kappa\gamma L}{1-\kappa\gamma L}\Big)\\ &+ 1 + 2i\frac{2-\gamma L}{\gamma L-1} = (1-\kappa) + 2i\Big[\frac{\kappa(2-\kappa\gamma L)}{1-\kappa\gamma L} + \frac{2-\gamma L}{\gamma L-1}\Big] \geq 0.\end{aligned}$$

2. **Case** $\kappa < 0$**:** it is sufficient to show

$$\frac{1 + 2i\frac{2-\gamma L}{\gamma L-1}}{(1-\gamma L)^{2i}} \geq 1 \geq -\kappa\frac{-1 + 2i(1 + \frac{1}{1-\kappa\gamma L})}{(1-\kappa\gamma L)^{2i}}.$$

The l.h.s. holds because $\gamma L \in (1, 2)$ and $(1-\gamma L)^{2i} < 1$. The r.h.s. is equivalent to:

$$(1-\kappa\gamma L)^{2i} \geq -\kappa\Big[-1 + 2i\Big(1 + \frac{1}{1-\kappa\gamma L}\Big)\Big].$$

Using $\gamma L \in (1, 2)$ and the second order Taylor approximation for $(1-\kappa)^{2i}$, we get the bounds:

$$\begin{aligned}(1-\kappa\gamma L)^{2i} \geq (1-\kappa)^{2i} &\geq 1 + 2i(-\kappa) + i(2i-1)\kappa^2\\ -\kappa\Big[-1 + 2i\Big(1 + \frac{1}{1-\kappa\gamma L}\Big)\Big] &\leq -\kappa\Big[-1 + 2i\left(1 + \frac{1}{1-\kappa}\right)\Big].\end{aligned}$$

Therefore, it is sufficient to show:

$$1 + 2i(-\kappa) + i(2i-1)\kappa^2 \geq -\kappa\Big[-1 + 2i\Big(1 + \frac{1}{1-\kappa}\Big)\Big],$$

which equivalently rewrites as the following inequality valid for all $i \geq 1$:

$$\kappa^2 i(i-1) + \Big(\kappa i + \frac{1}{1-\kappa}\Big)^2 + \frac{(1-\kappa)^3 - 1}{(1-\kappa)^2} \geq 0.$$

□

Lemma A.2 *Let $\kappa \in (-\infty, 1)$ and $\gamma L \in (1, \overline{\gamma L}_\infty(\kappa))$. Then there exists an index k such that $\delta_k(\gamma L, \kappa) < 0$ and $\delta_j(\gamma L, \kappa) < 0$ for all $j > k$.*

Proof The sequence δ_i is a difference of two monotone sequences: $\frac{2-\kappa\gamma L}{(1-\kappa\gamma L)^{2i}}$ and $\frac{2-\kappa\gamma L}{(1-\kappa\gamma L)^{2i}}$, respectively. The former is decreasing for $\kappa < 0$ (weakly convex functions) and increasing for $\kappa \geq 0$ (convex functions), whereas the later is always decreasing. In particular, $\delta_0(\gamma L, \kappa) > 0$. Because $(1-\gamma\mu)^2 \leq (1-\gamma L)^2$, to the limit it holds $\lim_{k\to\infty} \delta_k(\gamma L, \kappa) = -\infty$ and there exists some index k with $\delta_k(\gamma L, \kappa) < 0$. We prove by contradiction that $\delta_j(\gamma L, \kappa) < 0$ for all $j > k$; assume there exists some index $l > k$ such that $\delta_l(\gamma L, \kappa) \geq 0$. The following inequalities hold:

$$\begin{aligned} \delta_k < 0 &: \frac{2-\gamma\mu}{(1-\gamma\mu)^{2k}} < \frac{2-\gamma L}{(1-\gamma L)^{2k}} \\ \delta_l > 0 &: \frac{2-\gamma\mu}{(1-\gamma\mu)^{2l}} > \frac{2-\gamma L}{(1-\gamma L)^{2l}}. \end{aligned}$$

By multiplying the inequalities and performing the simplifications we get:

$$(1-\gamma\mu)^{2(k-l)} > (1-\gamma L)^{2(k-l)},$$

which implies, due to $(1-\gamma\mu)^2 > (1-\gamma L)^2$, that $k > l$, contradiction! □

Proof (**Proposition 2.12**)
1. Case $k = 0$ holds by definition since $\overline{\gamma L}_0(\kappa) = 1$. For any $k \geq 1$, by using Lemma A.2, condition $T_k(\overline{\gamma L}_k, \kappa) = 0$ implies $\delta_k(\gamma L, \kappa) \leq 0$ and hence $\delta_{k+1}(\gamma L, \kappa) < 0$. Furthermore, $T_{k+1}(\overline{\gamma L}_k, \kappa) \leq T_k(\overline{\gamma L}_k, \kappa) = 0$. Then it holds:

$$T_{k+1}(\overline{\gamma L}_k, \kappa) \leq T_k(\overline{\gamma L}_k, \kappa) = 0 = T_{k+1}(\overline{\gamma L}_{k+1}, \kappa).$$

The monotonicity of $T_k(\gamma L, \kappa)$ in stepsize γL (see Proposition 2.10) implies $\overline{\gamma L}_k(\kappa) < \overline{\gamma L}_{k+1}(\kappa)$.
2. Because stepsize thresholds sequence is monotone, it has a limit.

(a) **Case** $\kappa > 0$**.** By definition, $\overline{\gamma L}_k(\kappa) < \frac{2}{1+\kappa} < 2$ for all integers $k \geq 1$ and condition $T_k(\overline{\gamma L}_k(\kappa), \kappa) = 0$ implies:

$$\left(\frac{1-\overline{\gamma L}_k(\kappa)}{1-\kappa\overline{\gamma L}_k(\kappa)}\right)^{2k} = \kappa + (1-\kappa)(1-\overline{\gamma L}_k(\kappa))^{2k} \Leftrightarrow$$
$$\left|\frac{1-\overline{\gamma L}_k(\kappa)}{1-\kappa\overline{\gamma L}_k(\kappa)}\right| = \exp\left[\frac{\ln\left(\kappa + (1-\kappa)(1-\overline{\gamma L}_k(\kappa))^{2k}\right)}{2k}\right].$$

To the limit $k \to \infty$, the r.h.s. equals 1 because $(1-\overline{\gamma L}_k(\kappa))^{2k} \to 0$, hence $|1-\overline{\gamma L}_k(\kappa)| = |1-\kappa\overline{\gamma L}_k(\kappa)|$, which implies the solution $\overline{\gamma L}_\infty(\kappa) = \frac{2}{1+\kappa}$.

(b) **Case** $\kappa < 0$**.** By definition, $\overline{\gamma L}_k(\kappa) < 2$ for all integers $k \geq 1$ and condition $T_k(\overline{\gamma L}_k(\kappa), \kappa) = 0$ implies:

$$(1-\overline{\gamma L}_k(\kappa))^{-2k} = 1 + \frac{1-(1-\kappa\overline{\gamma L}_k(\kappa))^{-2k}}{-\kappa} \Leftrightarrow$$
$$\left|1-\overline{\gamma L}_k(\kappa)\right| = \exp\left[\frac{-\ln\left(1+\frac{1-(1-\kappa\overline{\gamma L}_k(\kappa))^{-2k}}{-\kappa}\right)}{2k}\right].$$

To the limit $k \to \infty$, the r.h.s. equals 1 because $(1-\kappa\overline{\gamma L}_k(\kappa))^{-2k} \to 0$ due to $\kappa < 0$. Hence, $\left|1-\overline{\gamma L}_k(\kappa)\right| = 1$, with the solution $\overline{\gamma L}_\infty(\kappa) = 2$.

(c) **Case** $\kappa = 0$**.** By definition, $\overline{\gamma L}_k(0) < 2$ and, similarly to the case $\kappa < 0$:

$$\left|1-\overline{\gamma L}_k(\kappa)\right| = \left[1+2k\overline{\gamma L}_k(\kappa)\right]^{\frac{-1}{2k}} = \exp\left[\frac{-\ln\left(1+2k\overline{\gamma L}_k(\kappa)\right)}{2k}\right].$$

To the limit $k \to \infty$, the r.h.s. equals 1, hence $\left|1-\overline{\gamma L}_k(\kappa)\right| = 1$, with the solution $\overline{\gamma L}_\infty(\kappa) = 2$.

3. Condition $\gamma L \in [\overline{\gamma L}_k(\kappa), \overline{\gamma L}_{k+1}(\kappa))$ implies $T_k(\overline{\gamma L}_k(\kappa), \kappa) \geq 0$ and $T_{k+1}(\overline{\gamma L}_k(\kappa), \kappa) < 0$. Because $\delta_i(\gamma L, \kappa) < 0$ for all $i \geq k$, we have (i) $T_i(\overline{\gamma L}_k, \kappa) > 0$ for $i = 1, \ldots, k-1$ (since k is uniquely defined) and (ii) $T_i(\overline{\gamma L}_k, \kappa) < 0$ for $i \geq k+1$. □

A.2 Proof of optimal constant stepsize for weakly convex functions

Proof (**Proposition 2.15**) Minimizing the upper bound from Theorem 2.4 is equivalent to maximizing the denominator $P_N(\gamma L, \gamma\mu)$ from (2.6). For a fixed curvature ratio $\kappa = \frac{\mu}{L}$, this reduces to maximizing the quantity $r(l) := l\, p(l, \kappa l)$, where $l \in (0, 2)$ is

the normalized stepsize γL and $p(l,u)$ is defined in (2.7):

$$r(l) = l\Big[2 - \tfrac{-\kappa l^2}{1-\kappa l-|1-l|}\Big] = \begin{cases} l\big(2 - \tfrac{\kappa l^2}{l-\kappa l}\big) & l \in (0,1]; \\ l\tfrac{(2-l)(2-\kappa l)}{2-l-\kappa l} & l \in [1, \overline{\gamma L}_1(\kappa)]. \end{cases}$$

Its first branch is concave in l and reaches its maximum for $l = 1$. Since this value is also feasible for the second branch, it means that the maximum belongs to the later expression and restrict the analysis to the interval $[1,2)$. One can check that $r(l)$ is strictly concave for $l \in [1,2)$. Moreover, $r'(1) = \frac{2}{(1-\kappa)^2} > 0$ and $r'(2) = -2(1-\frac{1}{\kappa}) < 0$, where

$$r'(l) = \frac{2}{(2-l-\kappa l)^2}\Big[-\kappa\big(1+\kappa\big)l^3 + \big[3\kappa + (1+\kappa)^2\big]l^2 - 4\big(1+\kappa\big)l + 4\Big]. \tag{A.1}$$

Hence, there exists an unique maximizer $l^* \in [1,2)$, further on denoted by $(\gamma L)^*(\kappa)$. The square bracket from (A.1) is exactly the expression from equation (2.14).

Moreover, the optimal stepsize $(\gamma L)^*(\kappa)$ strictly increases with κ, from $(\gamma L)^*(-\infty) = 1$ to $(\gamma L)^*(0) \nearrow 2$ and crosses the threshold $\overline{\gamma L}_1(\kappa)$ at some specific $\bar{\kappa}$ (see Figure 1), which results by taking $(\gamma L)^*(\kappa) = \overline{\gamma L}_1(\kappa)$ in (2.14). For longer stepsizes, some other transient and exponential terms in γ are involved (see expression of denominator (2.6)), therefore the solution $(\gamma L)^*(\kappa) > \overline{\gamma L}_1(\kappa)$ is only asymptotically optimal $(N \to \infty)$. □

A.3 Scheduled stepsizes

Proof **(Proposition 2.17)**

(i) One can check that $s_0 = \overline{\gamma L}_1(\kappa) > 1$, where $\overline{\gamma L}_1(\kappa)$ is given explicitly in Theorem 2.5. Further on, $s_k, s_+ \in (1, \frac{2}{1+[\kappa]_+})$, for $k \geq 0$. By definition of s_+:

$$0 = \frac{s_k}{2 - s_k(1+\kappa)} - \frac{s_+}{2 - s_+(1+\kappa)} + \frac{s_+(2-s_+)(2-\kappa s_+)}{2 - s_+(1+\kappa)}.$$

Since $\frac{s_+(2-s_+)(2-\kappa s_+)}{2-s_+(1+\kappa)} \geq 0$, it implies:

$$\frac{s_+}{2 - s_+(1+\kappa)} \leq \frac{s_k}{2 - s_k(1+\kappa)} \Leftrightarrow s_k \leq s_+.$$

After simplifications in (2.15), s_+ are the roots of the function:

$$h(s_+) = -\kappa[2 - s_k(1+\kappa)]s_+^3 + 2(1+\kappa)[2 - s_k(1+\kappa)]s_+^2 + 2[-3 + s_k(1+\kappa)]s_+ - 2s_k.$$

The demonstration of the existence of a unique root s_k^+ is divided into scenarios of weak and strong convexity.

(a) **Case** $\kappa \in (-\infty, 0]$**.** Then $s_k \in [1, 2)$, $s_0 = \overline{\gamma L}_1(0) = \frac{3}{2}$ and the following holds:

i. $h(1) = -(1-\kappa)(2-\kappa s_k) < 0$;
ii. $h(3/2) = \frac{-s_k(1-3\kappa)}{2} + \frac{9}{8}[2 - s_k(1+\kappa)] < 0$;
iii. $h(2) = 2(2 - s_k)$;
iv. $\frac{dh}{ds_+}(1) = (1+\kappa)(2-\kappa s_k) \geq 0$ for $\kappa \in [-1, 1)$;
v. $\frac{dh}{ds_+}(\frac{3}{2}) = 2 + [2 - s_k(1+\kappa)](2 - \frac{3}{4}\kappa) > 0$ for $\kappa \in (-\infty, 0]$;
vi. $\frac{dh}{ds_+}(2) = 2 + 4(1-\kappa)[2 - s_k(1+\kappa)] > 0$ for $\kappa \in (-\infty, 0]$;
vii. $\frac{d^2h(s_+)}{ds_+^2} = 2[2 - s_k(1+\kappa)][2 - \kappa - 3\kappa(s_+ - 1)] > 0$.

Therefore, h is strictly increasing and possesses a unique root $s_+ \in [\frac{3}{2}, 2)$.

(b) **Strongly convex case.** We restrict to $\kappa \in [0, 1)$ and $s_k \in [1, \frac{2}{1+\kappa})$. One can check:

i. $h(1) = -(1-\kappa)(2-\kappa s_k) < 0$;
ii. $h(\frac{2}{1+\kappa}) = \frac{2(1-\kappa)^2[2 - s_k(1+\kappa)]}{(1+\kappa)^3} > 0$;
iii. $\frac{dh}{ds_+}(1) = (1+\kappa)(2-\kappa s_k) > 0$;
iv. $\frac{dh}{ds_+}(\frac{2}{1+\kappa}) = 2(1+\kappa^2)[2 - s_k(1+\kappa)] + s_k\kappa(1+\kappa) + (1-\kappa)^2 > 0$;
v. $\frac{d^2h(s_+)}{ds_+^2} = 2[2 - s_k(1+\kappa)]\left[\frac{3\kappa[2 - s_k^+(1+\kappa)] + 2[\kappa + (1-\kappa)^2]}{1+\kappa}\right] > 0$.

Therefore, h is strictly increasing and possesses a unique root $s_+ \in [1, \frac{2}{1+\kappa})$.

(ii) Since $s_{k+1} = s_+$ is uniquely defined and $s_k \leq s_+$, the sequence is monotonically increasing.

(iii) Due to its monotonicity, the sequence $\{s_k(\kappa)\}_{k=-1}^{\infty}$ has a limit, denoted by s_∞. We prove by contradiction that $s_\infty = \frac{2}{1+[\kappa]_+}$. Assume the contrary, i.e., $s_\infty \neq \frac{2}{1+[\kappa]_+}$, which is the upper bound of the interval of roots, so that $s_\infty \in [1, \frac{2}{1+[\kappa]_+})$. By replacing it in the definition of (2.15), it holds $\frac{s_\infty(2-s_\infty)(2-\kappa s_\infty)}{2-s_\infty(1+\kappa)} = 0$, which implies one of the solutions $s_\infty \in \{0, 2, \frac{2}{\kappa}\}$, which are all outside of the interval $[1, \frac{2}{1+[\kappa]_+})$, hence the contradiction!

(iv) When $\kappa = 0$, we define $c_j := 2 - s_j(0), \forall j = \{-1, 0, \dots\}$. From the recurrence it results $2 - c_{j+1} - \frac{1}{c_{j+1}} + \frac{1}{c_j} = 0$. Since $\{s_j\}$ is monotonically increasing, with $s_0 = \frac{3}{2}$, we have that $c_0 = \frac{1}{2}$, $\{c_j\}$ is a decreasing sequence with $c_j \in (0, \frac{1}{2}]$, $\forall j \geq 0$, and $c_\infty = 0$. Thus, $\frac{1}{c_{j+1}} \geq \frac{1}{c_j} + \frac{3}{2}$, $\forall j \geq 0$. Telescoping for $j = 0, \dots, k-2$ we get $\frac{1}{c_{k-1}} \geq \frac{1}{c_0} + \frac{3}{2}(k-1)$, hence $c_{k-1} \leq \frac{2}{3k+1}$ and $s_{k-1} \geq 2 - \frac{1}{\frac{3}{2}k}$.

(v) When $\kappa \in (0, 1)$, we define $c_j := \frac{2}{1+\kappa} - s_j(\kappa), \forall j = \{-1, 0, \dots\}$. Since $\{s_j\}$ is monotonically increasing, then $\{c_j\}$ is a monotonically decreasing sequence with $c_\infty = 0$. Using $s_0 = \overline{\gamma L}_1(\kappa) = \frac{3}{1+\kappa+\sqrt{1-\kappa+\kappa^2}}$, one can check that $c_0 \leq \min\{\frac{2}{1+\kappa}(\frac{1-\kappa}{1+\kappa})^2, \frac{1}{2}\}$, thus $c_j \in (0, \frac{1}{2}]$, $\forall j \geq 0$. We prove $\frac{c_{j+1}}{c_j} \leq (\frac{1-\kappa}{1+\kappa})^2$,

$\forall j \geq 0$. By replacing in the recurrence of $s_j(\kappa)$ we get

$$\frac{(\frac{2}{1+\kappa} - c_{j+1})(\frac{2\kappa}{1+\kappa} + c_{j+1})(\frac{2}{1+\kappa} + \kappa c_{j+1})}{c_{j+1}(1+\kappa)} - \frac{2}{c_{j+1}(1+\kappa)^2} + \frac{2}{c_j(1+\kappa)^2} = 0.$$

After algebraic manipulations, this is equivalent with

$$\begin{aligned}\frac{c_{j+1}}{c_j} &= 1 - \frac{1+\kappa}{2}(\frac{2}{1+\kappa} - c_{j+1})(\frac{2\kappa}{1+\kappa} + c_{j+1})(\frac{2}{1+\kappa} + \kappa c_{j+1}) \\ &= (\frac{1-\kappa}{1+\kappa})^2 + \frac{\kappa(1+\kappa)c_{j+1}}{2}\Big[c_{j+1}^2 + 2c_{j+1}\frac{1+\kappa^3}{\kappa(1+\kappa)^2} - 4\frac{1+\kappa^3}{\kappa(1+\kappa)^3}\Big].\end{aligned}$$

For the expression within the square bracket, by exploiting its monotonicity with respect to c_{j+1} and evaluating it in $\frac{1}{2}$ we have that it is negative for any $c_{j+1} \in [0, \frac{1}{2}]$ and $\kappa \in (0, 1)$. Therefore, we get $\frac{c_{j+1}}{c_j} \leq (\frac{1-\kappa}{1+\kappa})^2$. By telescoping for $j = 0, \ldots, k-2$, this implies $c_{k-1} \leq c_0(\frac{1-\kappa}{1+\kappa})^{2(k-1)} \leq (\frac{1-\kappa}{1+\kappa})^{2k}$, hence $s_{k-1} \geq \frac{2}{1+\kappa} - (\frac{1-\kappa}{1+\kappa})^{2k}$.

□

A.4 Performance bounds for dynamic stepsizes

In this section we give the proof of Theorem 2.18, within which we also show the results from Corollary 2.19 (for convex functions), Corollary 2.20 (for strongly convex functions) and Corollary 2.21 (for weakly convex functions).

Proof (**Theorem 2.18**) By telescoping (N4SD) with stepsizes $\gamma_i L$:

$$\begin{aligned}f_0 - f_N \geq \frac{\gamma_{N-1}L}{2-\gamma_{N-1}L-\gamma_{N-1}\mu}\frac{\|g_N\|^2}{2L} + \frac{\gamma_0 L[(2-\gamma_0 L)(2-\gamma_0\mu)-1]}{2-\gamma_0 L-\gamma_0\mu}\frac{\|g_0\|^2}{2L} + \\ + \sum_{i=1}^{N-1}\frac{\|g_i\|^2}{2L}\Big[\frac{\gamma_{i-1}L}{2-\gamma_{i-1}L-\gamma_{i-1}\mu} + \frac{\gamma_i L[(2-\gamma_i L)(2-\gamma_i\mu)-1]}{2-\gamma_i L-\gamma_i\mu}\Big].\end{aligned} \tag{A.2}$$

Given that all weights are nonnegative (the fact which we prove below), the bound (2.16) is obtained by taking the minimum gradient norm:

$$f_0 - f_N \geq \min_{0\leq i\leq N}\left\{\frac{\|g_i\|^2}{2L}\right\}\sum_{i=0}^{N-1}\frac{\gamma_i L(2-\gamma_i L)(2-\gamma_i\mu)}{2-\gamma_i L-\gamma_i\mu}.$$

For $\kappa < 0$, we define $\tilde{N} := \arg\max_{0\leq i\leq N-1}\{s_i \leq (\gamma L)^*(\kappa)\}$.
Case (i): $\gamma_i L = s_i(\kappa)$ for $i = 0, \ldots, N-1$. This corresponds to the (strongly) convex case ($\kappa \in [0, 1)$) or weakly convex ($\kappa < 0$) with $N \leq \tilde{N}$. By the definition of $s_i(\kappa)$,

all weights but the one of $\|g_N\|$ vanish and the inequality (A.2) reduces to:

$$f_0 - f_N \geq \frac{s_{N-1}}{2 - s_{N-1}(1+\kappa)} \frac{\|g_N\|^2}{2L}, \tag{A.3}$$

which is of type (♣) and leads to the rate (2.17) from Corollary 2.20.
Moreover, in the convex case the expression of $s_k(0)$ can be computed analytically from (2.15) as the root belonging to $[1, 2)$ of:

$$\frac{s_k(0)(3 - 2s_k(0))}{2 - s_k(0)} + \frac{s_{k-1}(0)}{2 - s_{k-1}(0)} = 0.$$

This solution is $s_k(0) = \frac{3-2s_{k-1}(0)+\sqrt{9-4s_{k-1}(0)}}{2(2-s_{k-1}(0))}$, which equivalently rewrites as in the hypothesis of Corollary 2.19.
Case (ii): $\gamma_i L = \min\{s_i(\kappa)\,,\ (\gamma L)^*(\kappa)\}$ for $i = 0, \ldots, N-1$. This case corresponds to $\kappa < 0$ and $N > \tilde{N}$. Then inequality (A.2) becomes

$$f_0 - f_N \geq \left[\frac{s_{\tilde{N}}}{2 - s_{\tilde{N}}(1+\kappa)} + \frac{(\gamma L)^*[(2 - (\gamma L)^*)(2 - \kappa(\gamma L)^*) - 1]}{2 - (\gamma L)^*(1+\kappa)}\right] \frac{\|g_{\tilde{N}+1}\|^2}{2L} + \sum_{i=\tilde{N}+2}^{N-1} \left[\frac{(\gamma L)^*(2 - (\gamma L)^*)(2 - \kappa(\gamma L)^*)}{2 - (\gamma L)^*(1+\kappa)} \frac{\|g_i\|^2}{2L}\right] + \frac{(\gamma L)^*}{2 - (\gamma L)^*(1+\kappa)} \frac{\|g_N\|^2}{2L}. \tag{A.4}$$

It remains to prove the nonnegativity of $\|g_{\tilde{N}+1}\|^2$'s weight, namely:

$$\frac{s_{\tilde{N}}}{2 - s_{\tilde{N}}(1+\kappa)} + \frac{(\gamma L)^*[(2 - (\gamma L)^*)(2 - \kappa(\gamma L)^*) - 1]}{2 - (\gamma L)^*(1+\kappa)} \geq 0.$$

By definition of (2.15), it holds:

$$\frac{s_{\tilde{N}}}{2 - s_{\tilde{N}}(1+\kappa)} = -\frac{s_{\tilde{N}+1}[(2 - s_{\tilde{N}+1})(2 - \kappa s_{\tilde{N}+1}) - 1]}{2 - s_{\tilde{N}+1}(1+\kappa)},$$

with $s_{\tilde{N}+1} > (\gamma L)^*(\kappa)$. It remains to demonstrate

$$-\frac{s_{\tilde{N}+1}[(2 - s_{\tilde{N}+1})(2 - \kappa s_{\tilde{N}+1}) - 1]}{2 - s_{\tilde{N}+1}(1+\kappa)} \geq -\frac{(\gamma L)^*[(2 - (\gamma L)^*)(2 - \kappa(\gamma L)^*) - 1]}{2 - (\gamma L)^*(1+\kappa)},$$

which results from the monotone decreasing of the function $\sigma{:}[1, 2) \to \mathbb{R}$, $\sigma(t) := \frac{t[(2-t)(2-\kappa t)-1]}{2-t(1+\kappa)}$. This holds because $\frac{d\sigma(t)}{dt} = -\frac{2(1-\kappa t)(t-1)(1+2-t(1+\kappa)}{[2-t(1+\kappa)]^2} \leq 0$.
By taking the minimum gradient norm in (A.4), with $p(\gamma L, \gamma\mu)$ defined in (2.7):

$$f_0 - f_N \geq \min_{\tilde{N}+1 \leq i \leq N} \left\{\frac{\|g_i\|^2}{2L}\right\} \left[\frac{s_{\tilde{N}}}{2 - s_{\tilde{N}}(1+\kappa)} + p((\gamma L)^*, \kappa(\gamma L)^*)(N - \tilde{N} - 1)\right]. \tag{A.5}$$

Since sequence $\{s_k\}$ is increasing, the inequality $\frac{s_{\tilde{N}}}{2-s_{\tilde{N}}(1+\kappa)} \geq \frac{s_{N-1}}{2-s_{N-1}(1+\kappa)}$ is equivalent to $\tilde{N} \geq N-1$. Consequently, the merged r.h.s. of (A.3) and (A.5) can be expressed as a maximum:

$$\max\left\{\frac{s_{N-1}}{2-s_{N-1}(1+\kappa)}\,,\ \frac{s_{\tilde{N}}}{2-s_{\tilde{N}}(1+\kappa)} + p((\gamma L)^*, \kappa(\gamma L)^*)(N-\tilde{N}-1)\right\} = p((\gamma L)^*, \kappa(\gamma L)^*)N + \max_{0\leq k\leq N}\left\{\frac{s_{k-1}}{2-s_{k-1}(1+\kappa)} - p((\gamma L)^*, \kappa(\gamma L)^*)k\right\},$$

which is exactly as in (2.19) from Corollary 2.21. □

B Proofs on interpolability of smooth functions

B.1 Proof of interpolation conditions for smooth functions (Theorem 3.2)

Proof **(Theorem 3.2)** The proof follows the same steps outlined in [32, Theorem 4], with a minimal extension to accommodate $\mu < 0$. This extension impacts only the first step of the demonstration, known as minimal curvature subtraction, which involves shifting the lower curvature of the function to 0, effectively converting it into a smooth convex function. This adjustment remains valid for $\mu < 0$, as showed further on. By applying the *gradient inequality* for smooth convex functions: $h(x) \geq h(y) + \langle \nabla h(y), x-y\rangle\,,\ \forall x, y \in \mathbb{R}^d$ for $h = \frac{L}{2}\|\cdot\|^2 - f$ and $h = f - \frac{\mu}{2}\|\cdot\|^2$, we obtain:

$$\tfrac{\mu}{2}\|x-y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq \tfrac{L}{2}\|x-y\|^2 \quad \forall x, y \in \mathbb{R}^d. \quad \text{(B.1)}$$

Let $\bar{f} \in \mathcal{F}_{0,L-\mu}$, $\bar{f}(x) := f(x) - \frac{\mu}{2}\|x\|^2$, with $\nabla\bar{f}(x) = \nabla f(x) - \mu x$; by replacing f with $\bar{f}$ in the quadratic bounds inequality (B.1) it holds:

$$0 \leq \bar{f}(x) - \bar{f}(y) - \langle \nabla\bar{f}(y), x-y\rangle \leq \tfrac{L-\mu}{2}\|x-y\|^2.$$

Hence, the set $\{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ is $\mathcal{F}_{\mu,L}$-interpolable if and only if the set $\{(x_i, g_i - \mu x_i, f_i - \frac{\mu}{2}\|x_i\|^2)\}_{i\in\mathcal{I}}$ is $\mathcal{F}_{0,L-\mu}$-interpolable. Subsequently, the proof continues identically to [32, Theorem 4]. □

B.2 Proof of optimal point's characterization for smooth functions

Proof **(Lemma 5.5)** We extend the [15, Theorem 7] to smooth functions with any lower curvature μ, generalizing from the original convex ($\mu = 0$) and non-necessarily ($\mu = -L$) cases. Consider the function

$$Z(y) := \min_{\alpha\in\Delta_{\mathcal{I}}}\left\{\tfrac{L-\mu}{2}\Big\|y - \sum_{i\in\mathcal{I}}\alpha_i\big[x_i - \tfrac{1}{L-\mu}(g_i - \mu x_i)\big]\Big\|^2 + \sum_{i\in\mathcal{I}}\alpha_i\big(f_i - \tfrac{\mu}{2}\|x_i\|^2 - \tfrac{1}{2(L-\mu)}\|g_i - \mu x_i\|^2\big)\right\}$$

where $\Delta_{\mathcal{I}}$ is the $|\mathcal{I}|$-dimensional unit simplex:

$$\Delta_{\mathcal{I}} := \{\alpha \in \mathbb{R}^n : \sum_{i\in\mathcal{I}} \alpha_i = 1, \alpha_i \geq 0, \forall i \in \mathcal{I}\}.$$

The function Z is the primal interpolation function $W_{\mathcal{I}}^{C}$ as defined in [14, Definition 2.1]. One can check this by replacing in the definition $C \leftarrow \{0\}$, $L \leftarrow L - \mu$ and

$$\mathcal{T} \leftarrow \left\{ \left(x_i, g_i - \mu x_i, f_i - \tfrac{\mu}{2}\|x_i\|^2 \right) \right\}_{i\in\mathcal{I}}.$$

The set $\mathcal{T}$ is $\mathcal{F}_{0,L-\mu}$-interpolable. Hence, from [14, Theorem 1] it follows that Z is convex, with upper curvature $L - \mu$, and satisfies

$$Z(x_i) = f_i - \frac{\mu}{2}\|x_i\|^2 \quad \text{and} \quad \nabla Z(x_i) = g_i - \mu x_i.$$

Consider the function

$$\hat{W}(y) := Z(y) + \tfrac{\mu}{2}\|y\|^2,$$

which by curvature shifting belongs to the function class $\mathcal{F}_{\mu,L}$ and satisfies

$$\hat{W}(x_i) = f_i \quad \text{and} \quad \nabla\hat{W}(x_i) = g_i.$$

Using algebraic manipulations, $\hat{W}$ can be expressed as

$$\begin{aligned}\hat{W}(y) = \min_{\alpha\in\Delta_{\mathcal{I}}} \{ &\tfrac{L}{2}\big\|y - \sum_{i\in\mathcal{I}} \alpha_i \big(x_i - \tfrac{1}{L} g_i\big)\big\|^2 + \tfrac{L}{2}\tfrac{\mu}{L-\mu}\big\|\sum_{i\in\mathcal{I}} \alpha_i \big(x_i - \tfrac{1}{L} g_i\big)\big\|^2 + \\ &\sum_{i\in\mathcal{I}} \alpha_i \big(f_i - \tfrac{1}{2L}\|g_i\|^2 - \tfrac{L}{2}\tfrac{\mu}{L-\mu}\big\|x_i - \tfrac{1}{L} g_i\big\|^2\big)\}.\end{aligned} \tag{B.2}$$

Further, we *lower bound* the first squared norm by 0:

$$\hat{W}(y) \geq \min_{\alpha\in\Delta_{\mathcal{I}}} \{ \tfrac{L}{2}\tfrac{\mu}{L-\mu}\big\|\sum_{i\in\mathcal{I}} \alpha_i \big(x_i - \tfrac{1}{L} g_i\big)\big\|^2 + \sum_{i\in\mathcal{I}} \alpha_i \big(f_i - \tfrac{1}{2L}\|g_i\|^2 - \tfrac{L}{2}\tfrac{\mu}{L-\mu}\big\|x_i - \tfrac{1}{L} g_i\big\|^2\big)\}.$$

Applying Jensen inequality for the squared norm, which is a convex function, we get:

$$\big\|\sum_{i\in\mathcal{I}} \alpha_i \big(x_i - \tfrac{1}{L} g_i\big)\big\|^2 \leq \sum_{i\in\mathcal{I}} \alpha_i \big\|x_i - \tfrac{1}{L} g_i\big\|^2.$$

Then for $\mu \leq 0$ we can further lower bound the above inequality and obtain:

$$\hat{W}(y) \geq \min_{\alpha \in \Delta_{\mathcal{I}}} \left\{ \sum_{i \in \mathcal{I}} \alpha_i \left(f_i - \tfrac{1}{2L} \|g_i\|^2 \right) \right\} = \min_{i \in \mathcal{I}} \left\{ f_i - \tfrac{1}{2L} \|g_i\|^2 \right\} = f_{i_*} - \tfrac{1}{2L} \|g_{i_*}\|^2.$$

For the *upper bound*, in (B.2) we take $y := x_{i_*} - \frac{1}{L} g_{i_*}$ and $\alpha = e_{i_*}$ (the i_*-th unit vector):

$$\begin{aligned} \hat{W}\left(x_{i_*} - \tfrac{1}{L} g_{i_*}\right) &\leq \tfrac{L}{2} \tfrac{\mu}{L-\mu} \left\| x_{i_*} - \tfrac{1}{L} g_{i_*} \right\|^2 + f_{i_*} - \tfrac{1}{2L} \|g_{i_*}\|^2 - \tfrac{L}{2} \tfrac{\mu}{L-\mu} \left\| x_{i_*} - \tfrac{1}{L} g_{i_*} \right\|^2 \\ &= f_{i_*} - \tfrac{1}{2L} \|g_{i_*}\|^2. \end{aligned}$$

Therefore, from both lower and upper bounds it follows $\hat{W}\left(x_{i_*} - \frac{1}{L} g_{i_*}\right) = f_{i_*} - \frac{1}{2L} \|g_{i_*}\|^2$. □

C Tightness proofs (details)

C.1 Sublinear regimes with stepsizes $\gamma_i L \in (1, \overline{\gamma L}_1(\kappa)]$

Proof (**Proposition 5.9**) The equality case from Theorem 2.5's proof implies all gradient norms equal with each other as defined in (5.5). By replacing this expression in the equality conditions from Lemma 4.3, identity (5.6) results. Identity (5.7) is implied by the equality case in (N4SD):

$$f_i - f_{i+1} = \frac{\gamma_i L (2 - \gamma_i \mu)(2 - \gamma_i L)}{2 - \gamma_i L - \gamma_i \mu} \frac{U^2}{2L}.$$

The normalized inner product after one iteration is $c(\gamma_i L, \gamma_i \mu) \in (-1, 1)$, therefore $\theta(\gamma L, \gamma \mu)$ is well defined in (5.9). Moreover, recurrence (5.8) satisfies the necessary condition (5.6) since

$$\langle g_i, g_{i+1} \rangle = U^2 \cos\left((-1)^i \theta(\gamma_i L, \gamma_i \mu) \right) = U^2 c(\gamma_i L, \gamma_i \mu).$$

(5.10) holds because all gradient norms are equal and the function values decrease. □

Proof ***(Proposition*** 5.10 ***(cont.))*** We check that all interpolation conditions ($Q_{[i,j]}$) and ($Q_{[j,i]}$) hold $\forall (i, j) \in \{0, 1, \ldots, N\}$: (without loss of generality $j > i$)

$$Q_{[i,j]} : f_i - f_j - \gamma \langle g_j, \sum_{k=i}^{j-1} g_k \rangle \geq \tfrac{1}{2(L-\mu)} \left(\|g_i - g_j\|^2 + \gamma L \gamma \mu \| \sum_{k=i}^{j-1} g_k \|^2 - 2\gamma\mu \langle g_i - g_j, \sum_{k=i}^{j-1} g_k \rangle \right);$$

$$Q_{[j,i]} : f_j - f_i + \gamma \langle g_i, \sum_{k=i}^{j-1} g_k \rangle \geq \tfrac{1}{2(L-\mu)} \left(\|g_i - g_j\|^2 + \gamma L \gamma \mu \| \sum_{k=i}^{j-1} g_k \|^2 - 2\gamma\mu \langle g_i - g_j, \sum_{k=i}^{j-1} g_k \rangle \right).$$

Note the same expressions of the r.h.s. in both inequalities. Let $\Delta := f_0 - f_N$ and recall $c(\gamma L, \gamma\mu) = \frac{1+(1-\gamma L)(1-\gamma\mu)}{2-\gamma L-\gamma\mu}$. Then $1 + c = \frac{(2-\gamma L)(2-\gamma\mu)}{2-\gamma L-\gamma\mu}$ and

$$U^2 := \frac{\Delta}{\frac{1}{2L}\frac{\gamma L(2-\gamma L)(2-\gamma\mu)}{2-\gamma(L+\mu)}N} = \frac{2L\Delta}{(1+c)N}.$$

– **Case** $j - i$ **even.** Then $g_i = g_j$ and we use the identity:

$$\langle g_i \,,\, \sum_{k=i}^{j-1} g_k \rangle = \frac{U^2(1+c)(j-i)}{2}$$

to rewrite the interpolation inequalities as follows:

$$Q_{[i,j]}: \quad \frac{\Delta(j-i)}{N} - \frac{\gamma U^2(1+c)(j-i)}{2} \geq \frac{\gamma L\gamma\mu}{2(L-\mu)} \Big\| \sum_{k=i}^{j-1} g_k \Big\|^2;$$

$$Q_{[j,i]}: \quad \frac{\Delta(i-j)}{N} + \frac{\gamma U^2(1+c)(j-i)}{2} \geq \frac{\gamma L\gamma\mu}{2(L-\mu)} \Big\| \sum_{k=i}^{j-1} g_k \Big\|^2.$$

By replacing the definition of U^2, the l.h.s. in both inequalities equals zero, while the r.h.s. is non-positive because $\mu \leq 0$; therefore, both interpolation inequalities are valid for $i + j$ even.

– **Case** $j - i$ **odd.** Then $g_i \neq g_j$ and $\langle g_i, g_j \rangle = cU^2$, and we use the following identities:

$$\langle g_i \,,\, \sum_{k=i}^{j-1} g_k \rangle = \frac{U^2(j-i)(1+c)}{2} + \frac{U^2(1-c)}{2};$$

$$\langle g_j \,,\, \sum_{k=i}^{j-1} g_k \rangle = \frac{U^2(j-i)(1+c)}{2} - \frac{U^2(1-c)}{2};$$

$$\langle g_i - g_j, \sum_{k=i}^{j-1} g_k \rangle = U^2(1-c);$$

$$\Big\| \sum_{k=i}^{j-1} g_k \Big\|^2 = \frac{U^2}{2}\big[(1+c)(j-i)^2 + (1-c)\big];$$

$$\big\| g_i - g_j \big\|^2 = 2(1-c)U^2.$$

After substitutions, the r.h.s. and l.h.s. of both inequalities $Q_{[i,j]}$ and $Q_{[j,i]}$ are the same:

$$\text{r.h.s.}_{Q_{[i,j]}} = \text{r.h.s.}_{Q_{[j,i]}} = \frac{U^2}{4(L-\mu)} \Big[\gamma L\gamma\mu(1+c)(j-i)^2 + (1-c)\big(\gamma L\gamma\mu - 4\gamma\mu + 4\big)\Big]$$

and, respectively:

$$\text{l.h.s.}_{Q_{[i,j]}} : \frac{\Delta(j-i)}{N} - \gamma\Big\langle g_j, \sum_{k=i}^{j-1} g_k\Big\rangle = \frac{\gamma U^2(1-c)}{2};$$

$$\text{l.h.s.}_{Q_{[j,i]}} : \frac{\Delta(i-j)}{N} + \gamma\Big\langle g_i, \sum_{k=i}^{j-1} g_k\Big\rangle = \frac{\gamma U^2(1-c)}{2}.$$

It remains to prove (with $j-i = 1, 3, \dots$):

$$\frac{\gamma U^2(1-c)}{2} \geq \frac{U^2}{4(L-\mu)}\Big[\gamma L\gamma\mu(1+c)(j-i)^2 + (1-c)\big[(\gamma L)(\gamma\mu) - 4(\gamma\mu) + 4\big]\Big],$$

equivalent to

$$0 \geq \gamma L\gamma\mu(1+c)(j-i)^2 + (1-c)\big[\gamma L\gamma\mu - 2\gamma(L+\mu) + 4\big].$$

By using identity

$$(1-c)\big[\gamma L\gamma\mu - 2\gamma(L+\mu) + 4\big] = -\gamma L\gamma\mu(1+c),$$

the inequality to prove becomes:

$$0 \geq \gamma L\gamma\mu(1+c)\big[(j-i)^2 - 1\big].$$

This is true since $j-i \geq 1$, $\mu \leq 0$ and $c(\gamma L, \gamma\mu) \in [-1, 1]$; the equality case holds for $j = i+1$, i.e., the necessary conditions from equality in the distance-1 interpolation inequalities used to derive the upper bound.

We have shown that the proposed set of triplets satisfy the interpolation conditions for all pairs (i, j), hence, following Theorem 3.2, it is $\mathcal{F}_{\mu,L}$-interpolable. □

C.2 Proof of the three-dimensional worst-case example from Proposition 5.14

Preliminaries. We denote $\eta := 1 - \gamma\mu \in (1, \infty)$, $\rho := 1 - \gamma L \in (-1, 0)$, $\tilde{U} := \sqrt{\frac{(1-\rho^2)(\eta^2-1)}{\eta^2-\rho^2}}$, and rewrite the expressions defining the triplets as:

$$g_i = U\tilde{U}\begin{bmatrix} \frac{\rho^{i-\bar{N}}}{\sqrt{1-\rho^2}} \\ \frac{\eta^{i-\bar{N}}}{\sqrt{\eta^2-1}} \\ 0 \end{bmatrix} \quad \forall i = 0, 1, \ldots, \bar{N}+1;$$

$$g_{i+1} = \rho g_i + (\eta - \rho)\cos(\theta_{i+1})\begin{bmatrix} 0 & \\ & R(\theta_{i+1}) \end{bmatrix} g_i \quad \forall i = \bar{N}, \ldots, N. \tag{C.1}$$

$$f_i = \Delta\frac{N - i + \frac{(1-\rho)(1-\eta)}{\eta-\rho}\sum_{j=1}^{\bar{N}-i} T_j(\rho,\eta)}{N + \frac{(1-\rho)(1-\eta)}{\eta-\rho}\sum_{j=1}^{\bar{N}} T_j(\rho,\eta)} \quad i = 0, 1, \ldots, N. \tag{C.2}$$

The recurrence C.1 also works to define $g_{\bar{N}+1}$ due to $q_{\bar{N}+1} = 0$ and $\theta_{\bar{N}+1} = 0$. We consider

$$S_k := \sum_{p=0}^{k-(\bar{N}+2)} \rho^{2p} = \frac{1 - \rho^{2(k-(\bar{N}+1))}}{1-\rho^2}, \tag{C.3}$$

satisfying the properties $S_{k+1} = 1 + \rho^2 S_k$ and $S_\infty := \lim_{k\to\infty} S_k = \frac{1}{1-\rho^2}$. For each $k = \bar{N}+1, \bar{N}+2, \ldots, N$ we have that

$$q_k = \tan(\theta_k) = (-1)^{k-(\bar{N}+1)}\sqrt{(\eta^2-1)S_k}. \tag{C.4}$$

1. Proof of reaching the performance bound. We demonstrate that $\min_{0\le i\le\bar{N}} \|g_i\| = U$.
Note that:

$$\|g_i\|^2 = U^2\tilde{U}^2\left(\frac{\eta^{2(i-\bar{N})}}{\eta^2-1} + \frac{\rho^{2(i-\bar{N})}}{1-\rho^2}\right) \quad i = 0, \ldots, \bar{N}+1; \tag{C.5}$$

in particular, $\|g_{\bar{N}}\|^2 = \|g_{\bar{N}+1}\|^2 = U^2$. For the later, one can use the identity $\frac{\eta^2}{\eta^2-1} - \frac{\rho^2}{1-\rho^2} = \frac{1}{\eta^2-1} - \frac{1}{1-\rho^2}$, obtained by adding 1 to both ratios. We show that all gradients with indices $i \ge \bar{N}$ have norm U, namely $\|g_i\|^2 = U^2$ for any $i = \bar{N}, \ldots, N$.

Recurrence C.1 implies that the first component of g_i is always $U\tilde{U}\frac{\rho^{i-\bar{N}}}{\sqrt{1-\rho^2}}$, being a scaling with ρ of the first component of g_{i-1}. Let $v_i := \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} g_i$ be the vector of

the last two entries in g_i. Thus showing that $\|g_i\|^2 = U^2$ for $i \geq \bar{N}$ reduces to prove that

$$\|v_i\|^2 = U^2\Big(1 - \tilde{U}^2\frac{\rho^{2(i-\bar{N})}}{1-\rho^2}\Big), \ \forall i = \bar{N}, \bar{N}+1, \ldots, N. \tag{C.6}$$

We show this by induction. Firstly, it rewrites as

$$\|v_i\|^2 = U^2\tilde{U}^2\frac{1+q_{i+1}^2}{\eta^2-1}, \ \forall i = \bar{N}, \bar{N}+1, \ldots, N. \tag{C.7}$$

Cases $i = \{\bar{N}, \bar{N}+1\}$ result from the expressions of $g_{\bar{N}}$ and $g_{\bar{N}+1}$. Assuming that (C.6) holds for some index k, we show its validity for $k+1$. The recurrence of the gradients implies

$$v_{k+1} = \rho v_k + (\eta - \rho)\cos(\theta_{k+1})R(\theta_{k+1})v_k. \tag{C.8}$$

Taking the norm, expanding the squares, using the properties that the rotation matrix preserves the vector's norm and $\langle v_k, R(\theta_{k+1})v_k\rangle = \|v_k\|^2\cos(\theta_{k+1})$ yields

$$\|v_{k+1}\|^2 = \|v_k\|^2[\rho^2 + (\eta^2 - \rho^2)\cos^2(\theta_{k+1})].$$

Using $\cos^2(\theta_{k+1}) = \frac{1}{1+q_{k+1}^2}$ and the identity $\eta^2 + \rho^2 q_{k+1}^2 = 1 + q_{k+2}^2$ we get

$$\|v_{k+1}\|^2 = \|v_k\|^2\frac{1+q_{k+2}^2}{1+q_{k+1}^2},$$

where replacing $\|v_k\|$ using (C.7) yields the expression in (C.6) written for $k+1$.

Since $U^2 = \frac{2L\Delta}{\gamma L P_N(\gamma L, \gamma\mu)}$ and $f_0 - f_N = \Delta$, the triplets reach the performance bound with respect to $f(x_0) - f(x_N)$ from Proposition 5.14. To prove the bound with respect to $f(x_0) - f_*$, it remains to show that the required assumption in Lemma 5.6 holds, namely

$$N \in \underset{0\leq i\leq N}{\arg\min}\left\{f_i - \tfrac{1}{2L}\|g_i\|^2\right\}. \tag{C.9}$$

After replacing the expressions of gradient norms in (C.2) and several simplifications we get:

$$f_i - \frac{\|g_i\|^2}{2L} = f_N - \frac{U^2}{2L} + \frac{U^2}{2L}\frac{(1-\rho)(1+\rho)(1+\eta)}{\eta+\rho}\Big[N - \bar{N} + \sum_{j=1}^{\bar{N}-i}\eta^{-2j}\Big], \forall i = 0, \ldots, N,$$

where, with an abuse of notation, the sum is zero for $i \geq \bar{N}$. Its minimum is reached for $i = N$ and therefore identity (C.9) holds.

2. **The set of triplets $\mathcal{T}_{\mathcal{I}}$ is $\mathcal{F}_{\mu,L}$-interpolable.** We show that all interpolation conditions (3.1) are satisfied, namely for any i, j with $0 \leq i, j \leq N$, $i \neq j$, it holds that $Q_{i,j} \geq 0$, where

$$Q_{i,j} := f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{1}{2L}\|g_i - g_j\|^2 - \frac{\mu}{2L(L-\mu)}\|g_i - g_j - L(x_i - x_j)\|^2.$$

We divide the analysis into the cases: $i > j$ and $i < j$, showing that $Q_{i,j}$ rewrites as a sum of nonnegative contributions in each of them. The central challenge lies in finding an exact expression for the inner product $\langle g_{k+1} - \eta g_k \,,\, g_{i+1} - \rho g_i \rangle$ for all $i, k \geq \bar{N}$, which we obtain below in (C.15).

Preliminaries. For simplicity of notation and without loss of generality, we consider $U = 1$ since all gradients are multiplied by $U = \min_{0 \leq i \leq N} \|\nabla f(x_i)\|$.

Following (C.3), observe that $q_{\bar{N}+1} = 0$, $q_{\bar{N}+2} = -\sqrt{\eta^2 - 1}$ and $|q_k|$ is monotonically increasing towards $|q_\infty| = \sqrt{\frac{\eta^2-1}{1-\rho^2}}$. Consequently, $|\theta_k| = \arctan(|q_k|) \in (0, \frac{\pi}{2})$, $|\theta_k|$ monotonically increases towards $\arctan(|q_\infty|)$ and $\operatorname{sgn}(\theta_k) = (-1)^{k-(\bar{N}+1)}$. For all $k = 0, \ldots, \bar{N}+1$, the gradients read as $g_k = \tilde{U}[\frac{\rho^{k-\bar{N}}}{\sqrt{1-\rho^2}}, \frac{\eta^{k-\bar{N}}}{\sqrt{\eta^2-1}}, 0]^T$ and we have

$$\langle g_{k+1} - \eta g_k \,,\, g_{i+1} - \rho g_i \rangle = \langle [*, 0, 0]^T \,,\, [0, *, *]^T \rangle = 0, \ \forall k \leq \bar{N}, \ i = 0, \ldots, N-1, \tag{C.10}$$

where $*$ denotes for the entries of the three dimensional vectors multiplied by zero. Only the first component is $g_{k+1} - \eta g_k$ is nonzero (for $k \leq \bar{N}$), while the first component in $g_{i+1} - \rho g_i$ is always zero for any $i = 0, \ldots, N-1$.

Calculation of $\langle g_{k+1} - \eta g_k \,,\, g_{i+1} - \rho g_i \rangle$ for any $i, k \geq \bar{N}$. Identity (C.7) implies for all $k = \bar{N}, \bar{N}+1, \ldots, N-1$ that $\|v_k\| \cos(\theta_{k+1}) = \frac{\tilde{U}}{\sqrt{\eta^2-1}}$ and thus:

$$v_{k+1} - \rho v_k = (\eta - \rho)\frac{\tilde{U}}{\sqrt{\eta^2-1}} R(\theta_{k+1}) \frac{v_k}{\|v_k\|}; \tag{C.11}$$

$$\begin{aligned} v_{k+1} - \eta v_k &= -(\eta - \rho)\big[I_2 - \cos(\theta_{k+1}) R(\theta_{k+1})\big] v_k \\ &\overset{(C.7)}{=} -(\eta - \rho)\frac{\tilde{U}}{\sqrt{\eta^2-1}} q_{k+1} R\big(-\frac{\pi}{2}\big) R(\theta_{k+1}) \frac{v_k}{\|v_k\|}. \end{aligned} \tag{C.12}$$

Then we have

$$\langle g_{k+1} - \eta g_k \,,\, g_{i+1} - \rho g_i \rangle = \frac{-(\eta - \rho)^2 \tilde{U}^2}{\eta^2 - 1} q_{k+1} \cos(\beta_{k,i}), \tag{C.13}$$

where $\beta_{k,i} := \angle(R(-\frac{\pi}{2})R(\theta_{k+1})\frac{v_k}{\|v_k\|}, R(\theta_{i+1})\frac{v_i}{\|v_i\|})$. We normalize all vectors to transform the problem to maneuvering angles on the unit circle. Let $\phi(y)$ be the oriented angle of a two-dimensional vector y. We express the angles as follows:

$$\beta_{k,i} = \phi\big(R(\theta_{i+1})\frac{v_i}{\|v_i\|}\big) - \phi\big(R(-\frac{\pi}{2})R(\theta_{k+1})\big);$$

$$-\frac{\pi}{2} = \phi\big(R(-\frac{\pi}{2})R(\theta_{k+1})\frac{v_k}{\|v_k\|}\big) - \phi\big(R(\theta_{k+1})\frac{v_k}{\|v_k\|}\big);$$

$$\theta_{k+1} = \phi\big(R(\theta_{k+1})\frac{v_k}{\|v_k\|}\big) - \phi\big(\frac{v_k}{\|v_k\|}\big);$$

$$\theta_{i+1} = \phi\big(R(\theta_{i+1})\frac{v_i}{\|v_i\|}\big) - \phi\big(\frac{v_i}{\|v_i\|}\big).$$

Using these identities yields

$$\beta_{k,i} = \frac{\pi}{2} + \theta_{i+1} - \theta_{k+1} + \phi\big(\frac{v_i}{\|v_i\|}\big) - \phi\big(\frac{v_k}{\|v_k\|}\big).$$

For any $l \geq \bar{N}$, let $\delta_{l,l+1} := \phi\big(\frac{v_{l+1}}{\|v_{l+1}\|}\big) - \phi\big(\frac{v_l}{\|v_l\|}\big)$ be the oriented angle from $\frac{v_l}{\|v_l\|}$ to $\frac{v_{l+1}}{\|v_{l+1}\|}$ and $\epsilon_{l,l+1} := \theta_{l+2} - \theta_{l+1} + \delta_{l,l+1}$. We rewrite $\beta_{k,i}$ as

$$\beta_{k,i} = \frac{\pi}{2} + \sum_{l=\bar{N}}^{i-1}\epsilon_{l,l+1} - \sum_{l=\bar{N}}^{k-1}\epsilon_{l,l+1} = \begin{cases} \frac{\pi}{2} + \sum_{l=k}^{i-1}\epsilon_{l,l+1}, & \text{if } i \geq k; \\ \frac{\pi}{2} - \sum_{l=i}^{k-1}\epsilon_{l,l+1}, & \text{if } i < k. \end{cases}$$

We compute a closed form of $\sin(\epsilon_{l,l+1})$ with $l \geq \bar{N}$:

$$\sin(\epsilon_{l,l+1}) = \sin(\delta_{l,l+1})\cos(\theta_{l+2} - \theta_{l+1}) + \cos(\delta_{l,l+1})\sin(\theta_{l+2} - \theta_{l+1}).$$

Using that $q_l = \tan(\theta_l)$, we get:

$$\cos(\theta_{l+2} - \theta_{l+1}) = \frac{1 + q_{l+1}q_{l+2}}{\sqrt{(1+q_{l+2}^2)(1+q_{l+1}^2)}};$$
$$\sin(\theta_{l+2} - \theta_{l+1}) = \frac{q_{l+2} - q_{l+1}}{\sqrt{(1+q_{l+2}^2)(1+q_{l+1}^2)}}.$$

Since $\cos(\delta_{l,l+1}) = \langle\frac{v_l}{\|v_l\|}, \frac{v_{l+1}}{\|v_{l+1}\|}\rangle$ and $\sin(\delta_{l,l+1}) = \frac{v_l}{\|v_l\|} \times \frac{v_{l+1}}{\|v_{l+1}\|}$, we obtain

$$\cos(\delta_{l,l+1}) = \frac{\eta + \rho q_{l+1}^2}{\sqrt{(1+q_{l+2}^2)(1+q_{l+1}^2)}}; \ \sin(\delta_{l,l+1}) = \frac{(\eta - \rho)q_{l+1}}{\sqrt{(1+q_{l+2}^2)(1+q_{l+1}^2)}}.$$

Replacing in the expansion of $\sin(\epsilon_{l,l+1})$, after simplifications we get:

$$\sin(\epsilon_{l,l+1}) = \frac{\eta q_{l+2} - \rho q_{l+1}}{1 + q_{l+2}^2}.$$

Substituting q_{l+1} and q_{l+2} and amplifying by the conjugate square root yields:

$$\sin(\epsilon_{l,l+1}) = (-1)^{l-(\bar{N}+1)} \frac{\sqrt{\eta^2 - 1}}{\eta\sqrt{S_{l+2}} - \rho\sqrt{S_{l+1}}}.$$

For simplicity of notation, we define the angle α_l as

$$\alpha_l := \arcsin\left(\frac{\sqrt{\eta^2 - 1}}{\eta\sqrt{S_{l+2}} - \rho\sqrt{S_{l+1}}}\right), \ \forall l \geq \bar{N}. \tag{C.14}$$

Since $\rho < 0$, $\eta > 1$ and S_l is monotonically increasing with l, we have that α_l is monotonically decreasing with l, hence $\alpha_l \in [\alpha_\infty, \alpha_{\bar{N}}] \subset (0, \frac{\pi}{2})$, where $\alpha_{\bar{N}} = \arcsin\left(\sqrt{1 - \frac{1}{\eta^2}}\right)$ and $\alpha_\infty = \arcsin\left(\frac{\sqrt{(\eta^2-1)(1-\rho^2)}}{\eta-\rho}\right)$.
Using $q_{k+1} = \sqrt{(\eta^2 - 1)S_{k+1}}(-1)^{k-\bar{N}}$, after substituting everything into (C.13) we get

$$\langle g_{k+1} - \eta g_k \, , \, g_{i+1} - \rho g_i \rangle \, =$$

$$\frac{-(\eta-\rho)^2\tilde{U}^2}{\sqrt{\eta^2-1}}\sqrt{S_{k+1}} \begin{cases} \sin(\sum_{l=k}^{i-1}(-1)^{l-k}\alpha_l), & \text{if } \bar{N} \leq k \leq i \leq N-1; \\ \sin(\sum_{l=i}^{k-1}(-1)^{k-1-l}\alpha_l), & \text{if } \bar{N} \leq i < k \leq N-1. \end{cases} \tag{C.15}$$

All distance-1 interpolation inequalities $Q_{i,j}$ with $|i-j| = 1$ are involved in the proofs and thus hold with equality. This fact can also be checked by direct substitutions of our triplets and using that the inner product between consecutive gradients is known in closed form; for $i \geq \bar{N}$ we have ($\langle g_{\bar{N}}, g_{\bar{N}+1} \rangle = cU^2$), with $c = \frac{1+\eta\rho}{\eta+\rho}$ (as in (5.6)).
Case $i > j$**.** Let us consider a fixed $j \in \{0, 1, \ldots, N-1\}$. For each $i \in \{j+1, \ldots, N\}$, we define

$$\Delta Q_i := Q_{i+1,j} - Q_{i,j} - Q_{i+1,i}.$$

The distance-1 interpolation inequalities hold exactly, hence $Q_{i+1,i} = 0$ and $Q_{j+1,j} = 0$. Therefore to prove that $Q_{i,j} \geq 0$ for all $i > j$ it is sufficient to show that $\Delta Q_i \geq 0$, as $Q_{i,j}$ is obtained as a sum of positive increments.
Substituting the expressions of $Q_{i+1,j}$, $Q_{i,j}$ and $Q_{i+1,i}$ we subsequently obtain:

$$\Delta Q_i = \frac{1}{L-\mu}\left\langle g_i - g_j - \mu(x_i - x_j) \, , \, g_i - g_{i+1} - L(x_i - x_{i+1}) \right\rangle$$

$$= \frac{1}{L-\mu}\langle g_j - g_i - \gamma\mu \sum_{k=j}^{i-1} g_k \,,\; g_{i+1} - (1-\gamma L) g_i \rangle$$
$$= \frac{-\gamma}{\eta-\rho} \sum_{k=j}^{i-1} \langle g_{k+1} - \eta g_k \,,\; g_{i+1} - \rho g_i \rangle \,,$$

where in the second step we use the gradient descent iteration. For any $k \leq \bar{N}$, from (C.10) we have that $\langle g_{k+1} - \eta g_k \,,\; g_{i+1} - \rho g_i \rangle = 0$. Hence it is sufficient to focus on the range $j \geq \bar{N}+1$ (thus $i \geq \bar{N}+2$). Using (C.15) in the case $i \geq k$, the expression of ΔQ_i can be rewritten as

$$\Delta Q_i = \frac{\gamma(\eta-\rho)\tilde{U}^2}{\sqrt{\eta^2-1}} \sum_{k=j}^{i-1} \sqrt{S_{k+1}} \sin\Big(\sum_{l=k}^{i-1} (-1)^{l-k} \alpha_l \Big).$$

Because α_l decreases monotonically on $(0, \frac{\pi}{2})$ and the first term of the sum is always positive, then the sine of the sum is also positive, implying $\Delta Q_i \geq 0$.

Case $i < j$. Let us consider a fixed $j \in \{1, 2, \ldots, N\}$. For each $i \in \{0, 1, \ldots, j-1\}$, we define

$$\Delta Q_i := Q_{i,j} - Q_{i+1,j} - Q_{i,i+1}.$$

The distance-1 interpolation inequalities hold exactly, hence $Q_{i+1,i} = 0$ and $Q_{j+1,j} = 0$. Therefore to prove $Q_{i,j} \geq 0$ for all $i < j$ it is sufficient to show that $\Delta Q_i \geq 0$, since $Q_{i,j}$ is obtained as a sum of positive increments.
Similarly to the previous case, substituting the expressions of $Q_{i,j}$, $Q_{i+1,j}$ and $Q_{i,i+1}$, we subsequently obtain:

$$\Delta Q_i = \frac{1}{L-\mu} \left\langle g_{i+1} - g_j - \mu(x_{i+1} - x_j) \,,\; g_{i+1} - g_i - L(x_{i+1} - x_i) \right\rangle$$
$$= \frac{-\gamma}{\eta-\rho} \sum_{k=i+1}^{j-1} \langle g_{k+1} - \eta g_k \,,\; g_{i+1} - \rho g_i \rangle \,.$$

From (C.10) we have that $\Delta Q_i = 0$ if $j \leq \bar{N}+1$. Therefore, further on we assume $j \geq \bar{N}+2$. Moreover, we have the identity $g_{k+1} - \rho g_k = \eta^{\bar{N}-k}(g_{\bar{N}+1} - \rho g_{\bar{N}})$ for $k = 0, 1, \ldots, \bar{N}$, hence, using (C.10), ΔQ_i rewrites as

$$\Delta Q_i = \frac{-\gamma}{\eta-\rho} \eta^{\max\{0,\bar{N}-i\}} \sum_{k=\max\{i,\bar{N}\}+1}^{j-1} \left\langle g_{k+1} - \eta g_k \,,\; g_{\max\{i,\bar{N}\}+1} - \rho g_{\max\{i,\bar{N}\}} \right\rangle.$$

From (C.15) in the case $i < k$ we get

$$\Delta Q_i = \frac{\gamma(\eta-\rho)\tilde{U}^2\eta^{\max\{0,\bar{N}-i\}}}{\sqrt{\eta^2-1}} \sum_{k=\max\{i,\bar{N}\}+1}^{j-1} \sqrt{S_{k+1}} \sin\left(\sum_{l=\max\{i,\bar{N}\}}^{k-1} (-1)^{k-1-l}\alpha_l\right).$$

It is sufficient to check the positivity of the following expression for any $\bar{N} \le i < j$

$$V := \sum_{k=i+1}^{j-1} \sqrt{S_{k+1}} \sin\Big(\sum_{l=i}^{k-1}(-1)^{l-(k-1)}\alpha_l\Big).$$

Let $\Sigma_k := \sum_{l=i}^{k-1}(-1)^{l-(k-1)}\alpha_l$ and $Z_k := \sqrt{S_{k+1}}\sin(\Sigma_k)$. For any integer $p \ge 0$, one can check the following properties:

1. $\Sigma_{(i+1)+2p} + \Sigma_{(i+1)+2p+1} = \alpha_{(i+1)+2p}$;
2. $\Sigma_{(i+1)+2p}$ is a decreasing sequence;
3. $\Sigma_{(i+1)+2p} \ge \alpha_{(i+1)+2p}$;
4. $\Sigma_{(i+1)+2p} \in (0, \frac{\pi}{2})$.

Then we have

$$\begin{aligned} Z_{(i+1)+2p} &= \sqrt{S_{(i+1)+2p+1}}\sin(\Sigma_{(i+1)+2p}) > 0;\\ Z_{(i+1)+2p+1} &= -\sqrt{S_{(i+1)+2p+2}}\sin(\Sigma_{(i+1)+2p} - \alpha_{(i+1)+2p}) < 0. \end{aligned}$$

A sufficient condition for the positivity of V is $Z_{(i+1)+2p} + Z_{(i+1)+2p+1} \ge 0$ for any $p \ge 0$. This condition is equivalent to $\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}} > \frac{\sin(\Sigma_{(i+1)+2p}-\alpha_{(i+1)+2p})}{\sin(\Sigma_{(i+1)+2p})}$.
Using the monotonic decrease of the cotangent function on $(0, \frac{\pi}{2})$ and the property $\Sigma_{(i+1)+2p} \le \Sigma_{(i+1)} = \alpha_i$, we have that

$$\begin{aligned} \frac{\sin(\Sigma_{(i+1)+2p} - \alpha_{(i+1)+2p})}{\sin(\Sigma_{(i+1)+2p})} &= \sin(\alpha_{(i+1)+2p})(\cot(\alpha_{(i+1)+2p}) - \cot(\Sigma_{(i+1)+2p}))\\ &\le \sin(\alpha_{(i+1)+2p})(\cot(\alpha_{(i+1)+2p}) - \cot(\alpha_i)). \end{aligned}$$

From (C.14) we have $\cot(\alpha_l) = \frac{-\eta\rho\sqrt{S_{l+1}}+\sqrt{S_{l+2}}}{\sqrt{\eta^2-1}}$, for any $l \ge \bar{N}$. Then

$$\begin{aligned} &\sin(\alpha_{(i+1)+2p})(\cot(\alpha_{(i+1)+2p}) - \cot\alpha_i)\\ &= \frac{-\eta\rho\left(\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}} - \sqrt{\frac{S_{i+1}}{S_{(i+1)+2p+2}}}\right) + \left(1 - \sqrt{\frac{S_{i+2}}{S_{(i+1)+2p+2}}}\right)}{\eta - \rho\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}}}, \end{aligned}$$

where we amplified by $\frac{1}{\sqrt{S_{(i+1)+2p+2}}}$. Thus it is left to show that for each $p \geq 0$ it holds

$$\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}} \geq \frac{-\eta\rho\left(\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}} - \sqrt{\frac{S_{i+1}}{S_{(i+1)+2p+2}}}\right) + \left(1 - \sqrt{\frac{S_{i+2}}{S_{(i+1)+2p+2}}}\right)}{\eta - \rho\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}}},$$

which is equivalent to:

$$\eta(1+\rho) - \left(\rho + \frac{S_{(i+1)+2p+2}}{S_{(i+1)+2p+1}}\right)\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}} \geq \eta\rho\sqrt{\frac{S_{i+1}}{S_{(i+1)+2p+1}}} - \sqrt{\frac{S_{i+2}}{S_{(i+1)+2p+1}}}.$$

Using $\frac{S_{(i+1)+2p+2}}{S_{(i+1)+2p+1}} = \rho^2 + \frac{1}{S_{(i+1)+2p+1}}$, the inequality equivalently rewrites as:

$$(1+\rho)\left(\eta - \rho\sqrt{\frac{S_{(i+1)+2p+1}}{S_{(i+1)+2p+2}}}\right) + \frac{\frac{-1}{\sqrt{S_{(i+1)+2p+2}}} + \sqrt{S_{i+2}} - \eta\rho\sqrt{S_{i+1}}}{\sqrt{S_{(i+1)+2p+1}}} \geq 0.$$

The first term is positive because $\rho \in (-1, 0)$, whereas the second one due to $S_{i+2} \geq 1$ and $\frac{1}{\sqrt{S_{(i+1)+2p+2}}} < 1$. □

References

1. Abbaszadehpeivasti, H., de Klerk, E., Zamani, M.: The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions. Optimization Letters **16**(6), 1649–1661 (2022)
2. Altschuler, J., Parrilo, P.: Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, Nov (2024)
3. Altschuler, J., Parrilo, P.: Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. J. ACM **72**(2), 03 (2025)
4. Atenas, F., Sagastizábal, C., Silva, P.J.S., Solodov, M.: A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. SIAM Journal on Optimization **33**(1), 89–115 (2023)
5. Beck, A.: *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, (2017)
6. Bertsekas, D.: Nonlinear Programming. Athena Scientific, Athena scientific optimization and computation series (2016)
7. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, (2004)
8. Bubeck, S.: Convex optimization: Algorithms and complexity. Foundations and Trends in Machine Learning **8**, 231–357 (2015)
9. Carmon, Y., Duchi, J., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. Math. Program. **184**(1), 71–120 (2020)
10. Cartis, C., Gould, N., Toint, P.: On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. SIAM J. Optim. **20**(6), 2833–2852 (2010)
11. Das Gupta, S., Van Parys, B., Ryu, E.: Branch-and-bound performance estimation programming: a unified methodology for constructing optimal optimization methods. Mathematical Programming **204**(1), 641–641 (2024)

12. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
13. de Klerk, E., Glineur, F., Taylor, A.: On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. Optimization Letters **11**(7), 1185–1199 (2017)
14. Drori, Y.: The exact information-based complexity of smooth convex minimization. J. Complex. **39**, 1–16 (2017)
15. Drori, Y., Shamir, O.: The complexity of finding stationary points with stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 2658–2667. PMLR, (2020)
16. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Program. **145**(1–2), 451–482 (2014)
17. Evens, B., Pas, P., Latafat, P., Patrinos, P.: Convergence of the preconditioned proximal point method and Douglas–Rachford splitting in the absence of monotonicity. *Mathematical Programming*, Feb (2025)
18. Goujaud, B., Moucer, C., Glineur, F., Hendrickx, J., Taylor, A., Dieuleveut, A.: PEPit: computer-assisted worst-case analyses of first-order optimization methods in python. Math. Program. Comput. **16**(3), 337–367 (2024)
19. Grimmer, B.: Provably faster gradient descent via long steps. SIAM J. Optim. **34**(3), 2588–2608 (2024)
20. Grimmer, B., Shu, K., Wang, A.: Accelerated gradient descent via long steps. arXiv preprint arXiv:2309.09961, (2023)
21. Grimmer, B., Shu, K., Wang, A.: Accelerated objective gap and gradient norm convergence for gradient descent via long steps. INFORMS Journal on Optimization **7**(2), 156–169 (2025)
22. Kim, D., Fessler, J.: Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. J. Optim. Theory Appl. **188**(1), 192–219 (2021)
23. Kim, J.: A proof of the exact convergence rate of gradient descent. arXiv preprint arXiv:2412.04427, (2025)
24. Kim, J., Ozdaglar, A., Park, C., Ryu, E.K.: Time-reversed dissipation induces duality between minimizing gradient norm and function value. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, (2024). Curran Associates Inc
25. Lambert, D., Crouzeix, J.-P., Nguyen, V.H., Strodiot, J.-J.: Finite convex integration. Journal of Convex Analysis **11**(1), 131–146 (2004)
26. Nesterov, Y.: How to make the gradients small. *Optima 88*, (2012)
27. Nesterov, Y.: Lectures on Convex Optimization, 2nd edn. Springer Publishing Company, Incorporated (2018)
28. Rockafellar, R.T.: *Convex Analysis*. Princeton University Press, Princeton, New Jersey, (1970)
29. Taylor, A.: *Convex interpolation and performance estimation of first-order methods for convex optimization*. PhD thesis, UCLouvain - Université catholique de Louvain, Louvain-la-Neuve, (2017)
30. Taylor, A., Hendrickx, J., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization. SIAM J. Optim. **27**(3), 1283–1313 (2017)
31. Taylor, A., Hendrickx, J., Glineur, F.: Performance estimation toolbox (PESTO): Automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283, (2017)
32. Taylor, A., Hendrickx, J. T., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, (2017)
33. Taylor, A., Hendrickx, J., Glineur, F.: Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. J. Optim. Theory Appl. **178**(2), 455–476 (2018)
34. Teboulle, M., Vaisbourd, Y.: An elementary approach to tight worst case complexity analysis of gradient based methods. Math. Program. **201**(1–2), 63–96 (2022)
35. Themelis, A., Patrinos, P.: Douglas-Rachford splitting and admm for nonconvex optimization: Tight convergence results. SIAM J. Optim. **30**(1), 149–181 (2020)
36. Whitney, H.: Analytic extensions of differentiable functions defined in closed sets. *Hassler Whitney Collected Papers*, pages 228–254, (1934)
37. Whitney, H.: Differentiable functions defined in closed sets. i. *Transactions of the American Mathematical Society*, 36:369–387, (1934)
38. Zamani, M., Glineur, F.: Exact convergence rate of the last iterate in subgradient methods. SIAM Journal on Optimization, 35(3):2182–2201 (2025)

39. Zhang, Z., Jiang, R.: Accelerated gradient descent by concatenation of stepsize schedules. arXiv preprint arXiv:2410.12395, (2024)
40. Zihan Zhang, Jason Lee, Simon Du, and Yuxin Chen. Anytime acceleration of gradient descent. In Nika Haghtalab and Ankur Moitra, editors, Proceedings of Thirty Eighth Conference on Learning Theory, volume 291 of Proceedings of Machine Learning Research, pages 5991–6013. PMLR, (2025)