Anytime Acceleration of Gradient Descent

Zihan Zhang* U. Washington

Jason D. Lee† Princeton

Simon S. Du* U. Washington

Yuxin Chen‡ UPenn

November 26, 2024

Abstract

This work investigates stepsize-based acceleration of gradient descent with *anytime* convergence guarantees. For smooth (non-strongly) convex optimization, we propose a stepsize schedule that allows gradient descent to achieve convergence guarantees of $O(T^{-1.03})$ for any stopping time T, where the stepsize schedule is predetermined without prior knowledge of the stopping time. This result provides an affirmative answer to a COLT open problem (Kornowski and Shamir, 2024) regarding whether stepsize-based acceleration can yield anytime convergence rates of $o(T^{-1})$. We further extend our theory to yield anytime convergence guarantees of $\exp(-\Omega(T/\kappa^{0.97}))$ for smooth and strongly convex optimization, with κ being the condition number.

Contents

1 Introduction

Consider the standard problem of smooth convex optimization:

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad f(\boldsymbol{x}), \tag{1}$$

*Paul G. Allen School of Computer Science and Engineering, University of Washington; email: zihanz46@uw.edu, ssdu@cs.washington.edu.

†Department of Electrical and Computer Engineering, Princeton University; email: jasonlee@princeton.edu.

‡Department of Statistics and Data Science, University of Pennsylvania; email: yuxinc@wharton.upenn.edu.

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex (but not necessarily strongly convex). We assume without loss of generality that $f(\cdot)$ is 1-smooth (i.e., $\nabla f(\cdot)$ is 1-Lipschitz). In addition, we denote by $\boldsymbol{x}^*$ a minimizer of (1), and set $f^* = f(\boldsymbol{x}^*)$. Our focal point is the classical gradient descent (GD) algorithm:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \nabla f(\boldsymbol{x}_t), \qquad t \in \mathbb{N}, \tag{2}$$

where $\alpha_t > 0$ stands for the stepsize at iteration t, and $\boldsymbol{x}_0$ denotes the initialization.

Textbook gradient descent theory typically recommends a constant stepsize schedule $\alpha_t \equiv \alpha \in (0, 2)$, which ensures monotonicity of the objective value and guarantees that $f(\boldsymbol{x}_T) - f^* \leq O(1/T)$ for any stopping time T (Nesterov et al., 2018). Somewhat surprisingly, a recent strand of work (Altschuler, 2018; Altschuler and Parrilo, 2023a,b; Grimmer, 2024; Grimmer et al., 2023, 2024a; Rotaru et al., 2024; Teboulle and Vaisbourd, 2023) uncovered that adopting a time-varying stepsize schedule with occasional long steps can provably accelerate GD, achieving a convergence rate as fast as (Altschuler and Parrilo, 2023b; Grimmer et al., 2024a)

$$f(\boldsymbol{x}_T) - f^* \leq O\big(T^{-\log_2 \rho}\big) \quad \text{if } T = 2^k - 1 \text{ for some } k \in \mathbb{N}_+, \tag{3}$$

where $\rho := 1 + \sqrt{2}$ and $\log_2 \rho \approx 1.2716$. As a concrete example, this stepsize-based acceleration (3) is achievable via the so-called *silver stepsize schedule* (Altschuler and Parrilo, 2023b), which is constructed recursively and incorporates some large stepsizes far exceeding 2.

While occasional huge steps suffice in speeding up GD, the convergence guarantees (3) proven by Altschuler and Parrilo (2023b); Grimmer et al. (2023) only hold for exponentially increasing stopping times (i.e., $T = 2^k - 1$ for $k \in \mathbb{N}_+$). Given the non-monotonicity of $f(\boldsymbol{x}_t)$ in t due to the adoption of long steps, the intermediate points (i.e., those not corresponding to $t = 2^k - 1$) might incur significant sub-optimality gaps. In fact, it has been shown by Kornowski and Shamir (2024, Corollary 4) that the silver stepsize schedule cannot even guarantee $f(\boldsymbol{x}_t) - f^* \to 0$ at intermediate iterations.

To remedy this issue, Grimmer et al. (2024b); Zhang and Jiang (2024) proposed improved stepsize construction strategies that achieve $f(\boldsymbol{x}_T) - f^* \leq O(T^{-\log_2 \rho})$ for a prescribed stopping time T. One limitation of this approach is that it requires the stopping time T to be known in advance, as the stepsize schedule is designed based on the specific value of T. In practice, however, there is no shortage of applications where the stopping time is not predetermined and might vary during the execution of the algorithm. This gives rise to the following natural question, posed by Kornowski and Shamir (2024) at COLT 2024 as an open problem:

> ***Question:*** *Is there a stepsize schedule $\{\alpha_t\}_{t=1}^{\infty}$ that allows GD to achieve $f(\boldsymbol{x}_T) - f^* \leq o(1/T)$ for any stopping time $T \in \mathbb{N}$, where $\{\alpha_t\}_{t=1}^{\infty}$ is constructed without prior knowledge of T?*

In other words, this open problem asks whether it is feasible to achieve *anytime* convergence guarantees for GD that improve upon the textbook rate $O(1/T)$.

Overview of our results. In this work, we answer the above-mentioned open problem affirmatively. Our main finding is summarized below.

Theorem 1. *There exists a stepsize schedule $\{\alpha_t\}_{t=1}^{\infty}$, generated without knowing the stopping time, such that the gradient descent iterates (2) obey*[1]

$$f(\boldsymbol{x}_T) - f^* \leq O\bigg(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{T^{\vartheta}}\bigg) \qquad \textit{with } \vartheta = \frac{7 + \log_2 \rho}{8} > 1.03 \tag{4}$$

for an arbitrary stopping time $T \geq 1$.

To the best of our knowledge, our result provides the first stepsize schedule that provably accelerates gradient descent in an anytime fashion. The proposed stepsize schedule is inspired by, and constructed recursively based upon, the stepsize concatenation strategy recently proposed by Zhang and Jiang (2024) (see also Grimmer et al. (2024b)). While description of the precise stepsize schedule is postponed to Section 3.1, we immediately single out two features of our design: (i) the aggregate stepsize up to any time t is at least $\Omega(t^{1+\epsilon})$ for some constant $\epsilon > 0$, which often implies fast convergence; (2) it is guaranteed that the spikiness ratio is well-controlled in the sense that $\frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = o(t^{-\frac{1}{2}-\delta})$ for any $t > 0$, with some constant $\delta > 0$, which implies rapid convergence at those steps with spiky stepsizes.

[1] Throughout this paper, we use $\|\cdot\|$ to denote the ℓ_2 norm.

Other related work. In addition to the most relevant work described above, we mention in passing several other papers on gradient descent acceleration. Drori and Teboulle (2014) proposed the performance estimation problem (PEP) to identify tighter bounds on the worst-case GD performance under constant stepsize schedules. Taylor et al. (2017) put forward closed-form necessary and sufficient conditions for smooth (strongly) convex interpolation, offering a finite representation of these functions. Das Gupta et al. (2024) attempted to find the best possible worst-case convergence rate by solving the PEP via a branch-and-bound method. To improve the pre-constant in the $O(1/T)$ convergence rate, Teboulle and Vaisbourd (2023) proposed a dynamic bounded stepsize schedule, and Grimmer (2024) considered the periodic stepsize schedule. Both methods achieve highly non-trivial constant improvements. Additionally, Rotaru et al. (2024) studied the worst-case convergence rate for constant stepsize schedules for smooth non-convex functions, and established better convergence rates for weakly convex problems. There have also been a series of papers (Altschuler, 2018; Daccache et al., 2019; Eloi and Glineur, 2022) that computed the exact worst-case performance of GD for some fixed small iteration t. Noteworthily, most of the previous work focused on improving the worst-case convergence guarantees for a given stopping time T, instead of pursuing acceleration in an any-time fashion.

Paper organization. Section 2 introduces some basics about GD, as well as useful results from Zhang and Jiang (2024) concerning the so-called "primitive stepsize schedule." Construction of the proposed stepsize schedule and the proof of Theorem 1 provided in Section 3. In Section 4, we further extend our result to accommodate smooth and strongly convex optimization.

Notation. We also introduce a couple of notation to be used throughout. Denote by $\mathbf{1}$ the all-one vector with compatible dimension. Set

$$f_i = f(\boldsymbol{x}_i) \qquad \text{and} \qquad \boldsymbol{g}_i = \nabla f(\boldsymbol{x}_i) \tag{5}$$

for each iteration i. For a given stepsize schedule $\{\alpha_t\}_{t\geq 1}$, we set

$$A_n \coloneqq \sum_{i=1}^{n-1} \alpha_i \qquad \text{and} \qquad C_n \coloneqq \frac{A_n(A_n+1)}{2} \tag{6}$$

for any integer $n \geq 2$, where in the notation of A_n and C_n, we suppress the dependence on $\{\alpha_t\}_{t\geq 1}$ as long as it is clear from the context. Additionally, for an infinite sequence $\boldsymbol{r} = [r_j]_{j=1}^{\infty}$, we define

$$A_n(\boldsymbol{r}) = \sum_{i=1}^{n-1} r_i \qquad \text{and} \qquad C_n(\boldsymbol{r}) = \frac{A_n(\boldsymbol{r})(A_n(\boldsymbol{r})+1)}{2}. \tag{7}$$

In addition, we often use $\boldsymbol{\alpha}_{\ell:k}$ to indicate the stepsize subsequence $[\alpha_\ell, \ldots, \alpha_k]^\top$, and let $\alpha_i(\boldsymbol{s})$ denote the i-th stepsize in a stepsize sequence $\boldsymbol{s}$.

2 Preliminaries

Basic inequalities for smooth convex functions. Let us gather a set of elementary inequalities for a 1-smooth convex function $f(\cdot)$:

$$f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \leq 0, \tag{8a}$$

$$f^* - f_i + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \leq 0, \tag{8b}$$

$$f_i - f_j - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle + \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2 \leq 0, \tag{8c}$$

$$f_j - f_i - \langle \boldsymbol{g}_j, \boldsymbol{x}_j - \boldsymbol{x}_i \rangle + \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2 \leq 0, \tag{8d}$$

and for any $\boldsymbol{x}$ and $\alpha > 0$,

$$f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big) - f(\boldsymbol{x}) \\ \leq \alpha \big\langle \nabla f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big), \nabla f(\boldsymbol{x}) \big\rangle - \frac{1}{2} \big\| \nabla f(\boldsymbol{x}) - \nabla f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big) \big\|^2. \tag{8e}$$

See, e.g., Beck (2017) or Zhang and Jiang (2024, Section 2.1) for proofs of these well-known facts. In addition, given that $\alpha\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \alpha\|\boldsymbol{b}\|^2 + \alpha\langle \boldsymbol{a}-\boldsymbol{b}, \boldsymbol{b}\rangle \leq \alpha\|\boldsymbol{b}\|^2 + \frac{\alpha^2}{2}\|\boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|^2$ (a consequence of the Cauchy-Schwarz inequality), we can further upper bound (8e) by

$$f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big) - f(\boldsymbol{x}) \leq \frac{\alpha^2 + 2\alpha}{2} \|\nabla f(\boldsymbol{x})\|^2 \qquad \forall \alpha > 0 \text{ and } \boldsymbol{x}. \tag{8f}$$

Primitive stepsize schedule and concatenation. Next, we formalize the notion of "primitive stepsize schedule" as introduced in Zhang and Jiang (2024, Definition 3).

Definition 2 (Primitive stepsize schedule)**.** *A stepsize schedule $\boldsymbol{\alpha}_{1:k-1} = [\alpha_1, \ldots, \alpha_{k-1}] \in \mathbb{R}_+^{k-1}$ is said to be primitive if*

$$A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\ \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1} \alpha_i \left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right) \tag{9a}$$

$$\leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2, \tag{9b}$$

where we recall the definition of A_k and C_k in (6).

When $k = 1$, (9a) holds trivially, which means that the null sequence is a primitive stepsize schedule . As it turns out, two primitive stepsize schedules can be concatenated to form a longer primitive sequence, which forms the basis for the convergence guarantees in Zhang and Jiang (2024). The following lemma, derived by Zhang and Jiang (2024), makes precise this key property; for completeness, we provide a proof in Appendix B.1.

Lemma 3. *(Zhang and Jiang (2024, Theorem 3.1)) Consider a stepsize schedule $\{\alpha_t\}_{t\geq 1}$. Suppose that both $\boldsymbol{\alpha}_{1:\ell-1} = [\alpha_1, \ldots, \alpha_{\ell-1}]^\top$ and $\boldsymbol{\alpha}_{\ell+1:k-1} = [\alpha_{\ell+1}, \ldots, \alpha_{k-1}]^\top$ are primitive. Define the following function*

$$\varphi(x, y) \coloneqq \frac{-(x+y) + \sqrt{(x+y+2)^2 + 4(x+1)(y+1)}}{2}. \tag{10}$$

Then, $\boldsymbol{\alpha}_{1:k-1} = [\alpha_1, \ldots, \alpha_{k-1}]$ is also primitive if

$$\alpha_\ell = \varphi\big(\mathbf{1}^\top \boldsymbol{\alpha}_{1:\ell-1}, \mathbf{1}^\top \boldsymbol{\alpha}_{\ell+1:k-1}\big).$$

With Lemma 3 in mind, we find it convenient to introduce the concatenation function as follows: for any two nonnegative vectors $\boldsymbol{s}$ and $\boldsymbol{r}$, define

$$\mathsf{concat}(\boldsymbol{s}, \boldsymbol{r}) \coloneqq \big[\boldsymbol{s}^\top, \varphi(\mathbf{1}^\top\boldsymbol{s}, \mathbf{1}^\top\boldsymbol{r}), \boldsymbol{r}^\top\big]^\top. \tag{11}$$

As an immediate consequence, if we have available a collection of basic primitive sequences — denoted by $\{\boldsymbol{s}_i\}_{i\geq 1}$, then we can concatenate them as follows:

$$\widehat{\boldsymbol{s}}_0 = [\,], \tag{12a}$$

$$\widehat{\boldsymbol{s}}_i \leftarrow \mathsf{concat}(\widehat{\boldsymbol{s}}_{i-1}, \boldsymbol{s}_i), \qquad i = 1, 2, \ldots \tag{12b}$$

$$\widehat{\boldsymbol{s}} \leftarrow \lim_{i\to\infty} \widehat{\boldsymbol{s}}_i. \tag{12c}$$

The resulting $\widehat{\boldsymbol{s}}$ is well-defined and primitive, as asserted by the following lemma.

Lemma 4. *Suppose that each $\boldsymbol{s}_i$ ($i \geq 1$) is primitive. Then each $\widehat{\boldsymbol{s}}_i$ ($i \geq 1$) is primitive, and the infinite sequence $\widehat{\boldsymbol{s}}$ is well-defined and primitive.*

Proof. For each $i \geq 1$, $\widehat{\boldsymbol{s}}_{i-1}$ is always a prefix of $\mathsf{concat}(\widehat{\boldsymbol{s}}_{i-1}, \boldsymbol{s}_i) = \widehat{\boldsymbol{s}}_i$. As a result, for any $n \geq 1$, the n-th element of $\lim_{i\to\infty} \widehat{\boldsymbol{s}}_i$ exists, and hence $\widehat{\boldsymbol{s}}$ is well-defined.

Additionally, note that the null $\widehat{\boldsymbol{s}}_0$ is primitive. Assuming that $\widehat{\boldsymbol{s}}_{i-1}$ is primitive for some $i \geq 1$, we see from Lemma 3 that $\widehat{\boldsymbol{s}}_i = \mathsf{concat}(\widehat{\boldsymbol{s}}_{i-1}, \boldsymbol{s}_i)$ is also primitive. Therefore, an induction argument shows that $\widehat{\boldsymbol{s}}_i$ is primitive for every $i \geq 1$, and so is $\widehat{\boldsymbol{s}}$. $\square$

Silver stepsize schedule. We now introduce the silver stepsize schedule proposed by (Altschuler and Parrilo, 2023b).

Definition 5 (Silver stepsize schedule)**.** *Let $\overline{\boldsymbol{s}}_0 = [\,]$ be the null sequence, and set $\overline{\boldsymbol{s}}_i = \mathsf{concat}(\overline{\boldsymbol{s}}_{i-1}, \overline{\boldsymbol{s}}_{i-1})$ for each $i \geq 1$. Then $\overline{\boldsymbol{s}}_i$ is said to be the i-th order silver stepsize schedule, with the (limiting) silver stepsize schedule given by $\overline{\boldsymbol{s}} := \lim_{i\to\infty} \overline{\boldsymbol{s}}_i$.*

Given that $\overline{\boldsymbol{s}}_i$ is always a prefix of $\overline{\boldsymbol{s}}_{i+1} = \mathsf{concat}(\overline{\boldsymbol{s}}_i, \overline{\boldsymbol{s}}_i)$ for each $i \geq 0$, the limiting $\lim_{i\to\infty} \overline{\boldsymbol{s}}_i$ exists and hence $\overline{\boldsymbol{s}}$ is well-defined. Moreover, we single out the following properties about the silver stepsize schedule.

Lemma 6. *For each $i \geq 1$, $\overline{\boldsymbol{s}}_i$ is a primitive sequence with length $2^i - 1$. Moreover, it holds that*

$$\mathbf{1}^\top \overline{\boldsymbol{s}}_i = \rho^i - 1, \qquad i = 0, 1, \dots \tag{13}$$

where we recall that $\rho = \sqrt{2} + 1$.

Proof. First of all, Lemma 3 tells us that $\overline{\boldsymbol{s}}_{k+1} = \mathsf{concat}(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{s}}_k)$ is primitive as long as $\overline{\boldsymbol{s}}_k$ is primitive. Given that $\overline{\boldsymbol{s}}_0 = [\,]$ is also primitive, we can prove by induction that $\overline{\boldsymbol{s}}_i$ is primitive for every $i \geq 1$.

Next, we prove (13) by induction. To begin with, the claim (13) is trivial for $i = 0$. Now assuming that (13) holds for k, we have

$$\begin{aligned}
\mathbf{1}^\top \overline{\boldsymbol{s}}_{k+1} &= 2(\mathbf{1}^\top \overline{\boldsymbol{s}}_i) + \varphi(\mathbf{1}^\top \overline{\boldsymbol{s}}_k, \mathbf{1}^\top \overline{\boldsymbol{s}}_k) \\
&= 2(\rho^k - 1) + \left\{(\sqrt{2} - 1)(\rho^k - 1) + \sqrt{2}\right\} \\
&= \rho(\rho^k - 1) + \sqrt{2} = \rho^{k+1} - 1,
\end{aligned} \tag{14}$$

which justifies (13) for $i + 1$. This establishes (13) by induction. $\square$

3 Analysis

3.1 Construction of our stepsize schedule

Armed with the silver stepsize schedules $\{\overline{\boldsymbol{s}}_j\}_{j\geq 0}$ introduced in Definition 5 — which serve as basic primitive sequences — we can readily present the proposed stepsize schedule.

Take $c = 7$, and set

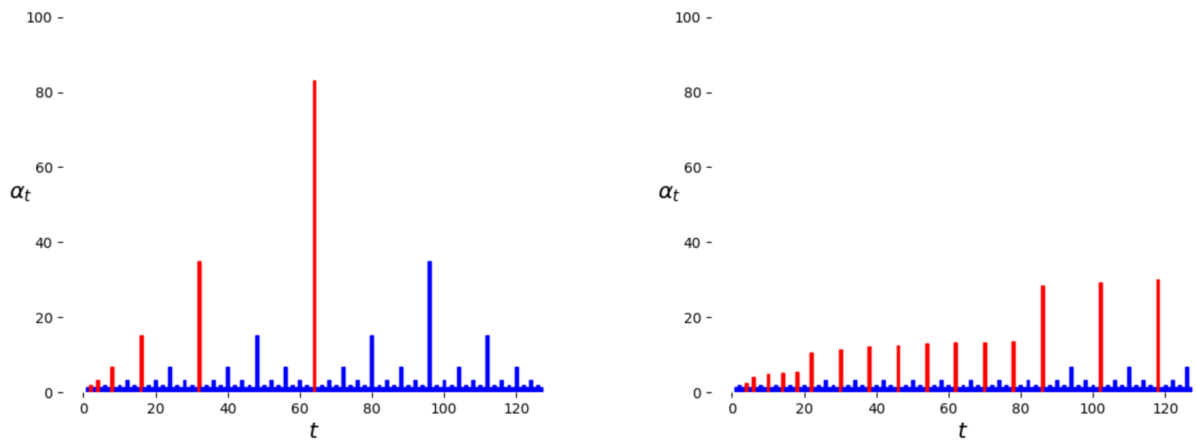
$$k_0 = M_0 = 0, \qquad k_i = 2^{ci}, \qquad \text{and} \qquad M_i = \sum_{j=1}^{i} k_i, \qquad i = 1, 2, \dots \tag{15}$$

With these parameters in place, our construction proceeds as follows:

- For each $j \geq 1$, set $\boldsymbol{s}_i = \overline{\boldsymbol{s}}_j$ for every i obeying $M_{j-1} < i \leq M_j$, where $\overline{\boldsymbol{s}}_j$ denotes the j-th order silver stepsize schedule in Definition 5. In other words, we repeat $\overline{\boldsymbol{s}}_j$ for k_j times for each $j \geq 1$, with k_j exponentially increasing in j.

- Generate the infinite stepsize sequence $\widehat{\boldsymbol{s}}$ through the concatenation procedure in (12).

Figure 1: **Left:** the first 128 steps of the silver stepsize schedule; **Right:** the first 128 steps of our stepsize schedule (with parameter c adjusted for better illustration). The **red bars** indicate the positions of the join steps. The number of join steps in the first t steps of the silver stepsize schedule is $\lfloor \log_2 t \rfloor$, whereas in our schedule, this number is roughly $\Omega(t^{7/8})$.



Throughout the rest of the paper, we denote by t_i the length of the i-th order subsequence $\widehat{\boldsymbol{s}}_i$, as constructed in (12).

We immediately single out an important property of the constructed stepsize schedule $\widehat{\boldsymbol{s}}$. The proof is postponed to Appendix A.

Lemma 7. *For any $t \geq 1$, it holds that*

$$A_{t+1}(\widehat{\boldsymbol{s}}) \geq \frac{1}{36} t^{\frac{c+\log_2 \rho}{c+1}}, \tag{16}$$

where $A_t(\widehat{\boldsymbol{s}})$ is defined in (7). Moreover, letting o_t denote the integer obeying $\sum_{j=1}^{o_t - 1} k_j 2^j < t \leq \sum_{j=1}^{o_t} k_j 2^j$, one has

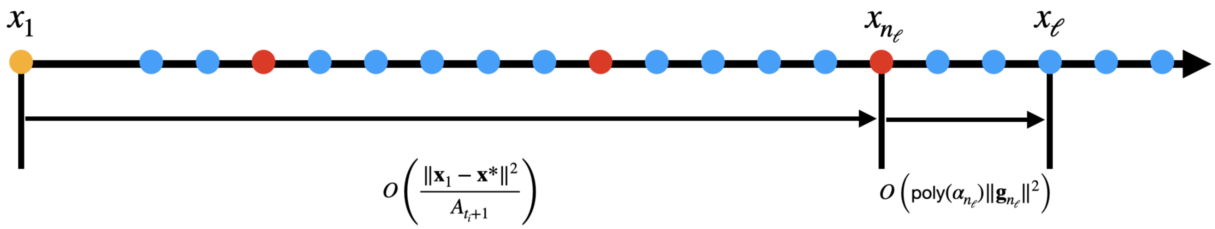
$$2^{o_t} \leq 2t^{\frac{1}{c+1}}. \tag{17}$$

3.2 A glimpse of high-level ideas

Let us take a moment to briefly point out two key aspects underlying our design and analysis of the stepsize schedule.

Stepsize concatenation via suitable join steps. As proven recently by Grimmer et al. (2024b); Zhang and Jiang (2024), certain desirable stepsize schedules with different lengths can be concatenated — with a properly chosen join stepsize — into a longer stepsize schedule while ensuring fast convergence at the last step, which motivates our design. To be more concrete, a desirable stepsize schedule of this kind is the primitive stepsize schedule, and it has been shown that a primitive stepsize schedule with length t enjoys the convergence rate of $O\big(\frac{1}{\sum_{i:i<t} \alpha_i}\big)$ at the last step (Zhang and Jiang, 2024). As a result, if we recursively prolong the stepsize schedule by concatenating the current one with another primitive stepsize schedule, then the $O\big(\frac{1}{\sum_{i<t} \alpha_i}\big)$ convergence rate continues to hold at the last step. Notably, every concatenation operation requires inserting a join stepsize in the middle, which we illustrate in Figure 1. As it turns out, there is a trade-off between the aggregate stepsize $\sum_{i:i<t} \alpha_i$ and the number of join steps, making it crucial to choose a proper number of join steps. Fortunately, there exists some simple stepsize schedule with $\Omega(t^{1-\epsilon_1})$ join steps and an aggregate stepsize $\Omega(t^{1+\epsilon_2})$ for some proper constants $\epsilon_1, \epsilon_2 > 0$, which enables a convergence rate of $o(t^{-1})$ at each join step.

Figure 2: An illustration of our analysis strategy to bound $f_\ell - f^*$ for an intermediate step ℓ. Here, the **yellow point** indicates the initial step, whereas the **red points** indicate the join steps. Here, n_ℓ indicates the largest join step below ℓ.



Controlling the norm of gradients. While the above-mentioned concatenation strategy guarantees fast convergence at each join step, we still need to examine the convergence properties at intermediate steps (i.e., the ones between two adjacent join steps). Consider, for concreteness, iteration ℓ, and denote by n_ℓ the closest join iteration below ℓ; see Figure 3.2 for an illustration. A common strategy to bound the difference $f_\ell - f_{n_\ell}$ of the associated objective values is to control the norm of the weighted gradients $\alpha_i^2\|\boldsymbol{g}_i\|^2$ for every $i \in [n_\ell, \ell]$, which arises from the smoothness and convexity of f. A key part of our analysis thus boils down to bounding each $\alpha_i^2\|\boldsymbol{g}_i\|^2$ using the corresponding weighted gradient at the join step n_ℓ, for which the silver stepsize schedule enjoys some favorable property that enables effective control of $\alpha_i^2\|\boldsymbol{g}_i\|^2$ in this manner.

3.3 Key lemmas

Before proceeding to the proof of our main theorem, we single out a couple of key lemmas concerning the primitive stepsize schedule — and in particular, the silver stepsize schedule — that play an important role in our subsequent analysis.

The first lemma below singles out an important property of a primitive stepsize schedules, to be specified by (18).

Lemma 8. *Suppose $\boldsymbol{s} = \boldsymbol{\alpha}_{1:k-1}$ is a primitive stepsize schedule. Then for any fixed $\boldsymbol{x}_0$ with gradient $\boldsymbol{g}_0$, it holds that*

$$A_k(f_k - f_0) + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2 + C_k\|\boldsymbol{g}_k\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \sum_{i=1}^{k-1}\alpha_i\langle \boldsymbol{g}_i, \boldsymbol{g}_0\rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2. \tag{18}$$

Proof. From Definition 2 of the primitive stepsize schedule, we obtain

$$\begin{aligned} & A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\ \le \quad & \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right). \end{aligned} \tag{19}$$

Also, the basic properties about smooth convex functions (cf. (8)) give

$$\sum_{i=1}^{k-1}\alpha_i\left(f_i - f_0 - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_0\rangle + \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_0\|^2\right) \le 0, \tag{20}$$

which further implies that

$$\begin{aligned} & \sum_{i=1}^{k-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right) \\ & \le \sum_{i=1}^{k-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right) - \sum_{i=1}^{k-1}\alpha_i\left(f_i - f_0 - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_0\rangle + \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_0\|^2\right) \end{aligned}$$

$$= \sum_{i=1}^{k-1} \alpha_i \left(f_0 - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{1}{2}\|\boldsymbol{g}_0\|^2 \right). \tag{21}$$

Substituting (21) into (19) and using the fact that $\sum_{i=1}^{k-1} \alpha_i \boldsymbol{g}_i = \boldsymbol{x}_1 - \boldsymbol{x}_k$, we can derive

$$\begin{aligned}
&A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\
&\le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1} \alpha_i \left(f_0 - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{1}{2}\|\boldsymbol{g}_0\|^2 \right) \\
&= \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + A_k(f_0 - f^*) - \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2 \\
&= \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + A_k(f_0 - f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_k, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2. \qquad (22)
\end{aligned}$$

To continue, we make the observation that

$$\begin{aligned}
\frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 - \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \langle \boldsymbol{x}_1 - \boldsymbol{x}_k, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle &= \frac{1}{2}\|\boldsymbol{x}_1\|^2 - \frac{1}{2}\|\boldsymbol{x}_k\|^2 - \langle \boldsymbol{x}_1, \boldsymbol{x}_0 \rangle + \langle \boldsymbol{x}_k, \boldsymbol{x}_0 \rangle \\
&= \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 - \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2,
\end{aligned}$$

which combined with (22) yields

$$A_k(f_k - f_0) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2$$

as claimed. □

Furthermore, the result in Lemma 8 allows us to control the gradient norm at the last step, provided that a primitive stepsize schedule is adopted.

Lemma 9. *Assume $\boldsymbol{s} = \boldsymbol{\alpha}_{1:k-1}$ is a primitive stepsize schedule. Assume $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha_0 \boldsymbol{g}_0$. Then one has*

$$C_k\|\boldsymbol{g}_k\|^2 \le \left(\frac{\alpha_0^2}{2} + \frac{(A_k+1)^2}{2} - \alpha_0 - \frac{A_k}{2} \right) \|\boldsymbol{g}_0\|^2. \tag{23}$$

Proof. Because $\boldsymbol{\alpha}_{1:k-1}$ is a primitive stepsize schedule, it follows from Lemma 8 that

$$A_k(f_k - f_0) + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2 + C_k\|\boldsymbol{g}_k\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2.$$

We also make note of the following basic facts:

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha_0 \boldsymbol{g}_0; \tag{24a}$$

$$\sum_{i=1}^{k-1} \alpha_i \boldsymbol{g}_i = \boldsymbol{x}_1 - \boldsymbol{x}_k = \boldsymbol{x}_0 - \boldsymbol{x}_k - \alpha_0 \boldsymbol{g}_0; \tag{24b}$$

$$f_0 - f_k \le \langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_k \rangle - \frac{1}{2}\|\boldsymbol{g}_0 - \boldsymbol{g}_k\|^2; \tag{24c}$$

$$(A_k + 1) \langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_k \rangle \le \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2 + \frac{(A_k+1)^2}{2}\|\boldsymbol{g}_0\|^2. \tag{24d}$$

Putting the above inequalities together, we arrive at

$$C_k\|\boldsymbol{g}_k\|^2 \le \left(\frac{\alpha_0^2}{2} + \frac{(A_k+1)^2}{2} - \alpha_0 - \frac{A_k}{2} \right) \|\boldsymbol{g}_0\|^2 \tag{25}$$

as claimed. □

Additionally, the following lemma enables effective control of the gradient norms in all intermediate steps.

Lemma 10. *Fix $i \geq 1$ and $\alpha \geq 0$, and let $k = 2^i$. Denote by $\overline{\boldsymbol{s}}_i = [\alpha_1, \ldots, \alpha_{k-1}]^\top$ the first $k-1$ stepsizes of the i-th order silver stepsize schedule $\overline{\boldsymbol{s}}_i$. Fix $\boldsymbol{x}_0$, set $\alpha_0 = \alpha$, and let $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha_0 \boldsymbol{g}_0$. If $\alpha \geq \max\{(\sqrt{2}-1)A_k, 1\}$, then one has*

$$\alpha_\ell^2 \|\boldsymbol{g}_\ell\|^2 \leq 16^i \alpha^2 \|\boldsymbol{g}_0\|^2 \tag{26a}$$

$$f_\ell - f_0 \leq \frac{3}{2} \cdot (32^i \alpha^2 \|\boldsymbol{g}_0\|^2) \tag{26b}$$

for any $1 \leq \ell \leq k-1$.

Proof. We shall prove the first claim of this lemma by induction. Regarding the base case with $i = 1$, from the fact that $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha \boldsymbol{g}_0$ and $\alpha_1 = \sqrt{2}$. The convexity and 1-smoothness of f give

$$\|\boldsymbol{g}_1 - \boldsymbol{g}_0\|^2 \leq \langle \boldsymbol{g}_0 - \boldsymbol{g}_1, \boldsymbol{x}_0 - \boldsymbol{x}_1 \rangle = \alpha \langle \boldsymbol{g}_0 - \boldsymbol{g}_1, \boldsymbol{g}_0 \rangle.$$

Rearrange terms to yield

$$\begin{aligned}
&\|\boldsymbol{g}_1\|^2 \leq (\alpha - 1)\|\boldsymbol{g}_0\|^2 - (\alpha - 2)\langle \boldsymbol{g}_0, \boldsymbol{g}_1 \rangle \leq (\alpha - 1)\|\boldsymbol{g}_0\|^2 + \frac{(\alpha-2)^2}{2}\|\boldsymbol{g}_0\|^2 + \frac{1}{2}\|\boldsymbol{g}_1\|^2 \\
\Longrightarrow \qquad & \|\boldsymbol{g}_1\|^2 \leq \big((\alpha-2)^2 + 2\alpha - 2\big)\|\boldsymbol{g}_0\|^2 \leq 16\alpha^2 \|\boldsymbol{g}_0\|^2.
\end{aligned}$$

Next, assume the claim (26a) holds for $i-1$ for some $i \geq 2$, then it suffices to prove it for $\ell \geq 2^{i-1}$. Let $\ell^* = 2^{i-1}$. Because $\overline{\boldsymbol{s}}_{i-1} = \boldsymbol{\alpha}_{1:\ell^*-1}$ is a primitive stepsize schedule, it follows from Lemma 9 that

$$C_{\ell^*} \|\boldsymbol{g}_{\ell^*}\|^2 \leq \left(\frac{\alpha^2}{2} + \frac{(A_{\ell^*}+1)^2}{2} - \alpha - \frac{A_{\ell^*}}{2} \right) \|\boldsymbol{g}_0\|^2. \tag{27}$$

Recall that $\alpha_{\ell^*} = \varphi(A_{\ell^*}, A_{\ell^*}) = (\sqrt{2}-1)A_{\ell^*} + \sqrt{2}$ and $A_{\ell^*} \geq \sqrt{2}$ (a consequence of Lemma 7). Recognizing that $\frac{\alpha_{\ell^*}}{A_{\ell^*}} = \frac{(\sqrt{2}-1)A_{\ell^*}+\sqrt{2}}{A_{\ell^*}} \leq \sqrt{2}$, $C_{\ell^*} = \frac{A_\ell^*(A_\ell^*+1)}{2}$ and $\alpha \geq \min\{(\sqrt{2}-1)A_k, 1\}$, we can deduce that

$$\begin{aligned}
\alpha_{\ell^*}^2 \|\boldsymbol{g}_{\ell^*}\|^2 &= \frac{\alpha_{\ell^*}^2}{C_{\ell^*}} \cdot C_{\ell^*} \|\boldsymbol{g}_{\ell^*}\|^2 \\
&\leq 2\frac{\alpha_{\ell^*}^2}{A_{\ell^*}(A_{\ell^*}+1)} \cdot \left(\frac{1}{2} + \frac{A_{\ell^*}^2 + A_{\ell^*}}{2\alpha^2} \right) \alpha^2 \|\boldsymbol{g}_0\|^2 \\
&\leq \left(2 + \frac{\alpha_{\ell^*}^2}{\alpha^2} \right) \alpha^2 \|\boldsymbol{g}_0\|^2 \\
&= \left(2 + \frac{\alpha_{\ell^*}^2}{A_{\ell^*}^2} \cdot \frac{A_{\ell^*}^2}{\alpha^2} \right) \alpha^2 \|\boldsymbol{g}_0\|^2 \\
&\leq \big(2 + (2+\sqrt{2})^2\big)\alpha^2 \|\boldsymbol{g}_0\|^2 \\
&\leq 16\alpha^2 \|\boldsymbol{g}_0\|^2.
\end{aligned} \tag{28}$$

Applying induction over the $(i-1)$-th order silver stepsize schedule $\boldsymbol{\alpha}_{2^{i-1}+1:2^i-1}$ with $\alpha = \alpha_{2^{i-1}} = (\sqrt{2}-1)A_{\ell^*} + \sqrt{2} \geq \min\{(\sqrt{2}-1)A_{\ell^*}, 1\}$, we can show that

$$\alpha_\ell^2 \boldsymbol{g}_\ell^2 \leq 16^{i-1} \alpha_{\ell^*}^2 \|\boldsymbol{g}_{\ell^*}\|^2 \leq 16^i \alpha^2 \|\boldsymbol{g}_0\|^2, \qquad \ell \in [2^{i-1}, 2^i - 1]. \tag{29}$$

We now turn to the second claim of this lemma. From (8f) and the fact that each stepsize in $\overline{\boldsymbol{s}}_i$ is at least 1 for all $i \geq 1$ — which holds since $\varphi(x, y) > 1$ for all $x, y \geq 0$ — we see that

$$\begin{aligned}
f_\ell - f_0 &= \sum_{j=0}^{\ell-1} (f_{j+1} - f_j) \leq \sum_{j=0}^{\ell-1} \frac{\alpha_j^2 + 2\alpha_j}{2} \|\boldsymbol{g}_i\|^2 \\
&\leq \frac{3}{2} \left(\sum_{j=0}^{\ell-1} \alpha_i^2 \|\boldsymbol{g}_i\|^2 \right) \leq \frac{3}{2} \cdot \big(32^i \alpha^2 \|\boldsymbol{g}_0\|^2\big),
\end{aligned}$$

which concludes the proof. $\square$

3.4 Proof of Theorem 1

We are now positioned to prove our main result in Theorem 1, based on the stepsize schedule $\widehat{\boldsymbol{s}} = [\alpha_i]_{i=1}^{\infty}$ constructed in Section 3.1. Let us remind the readers of several notation below.

- t_i: the length of the i-th subsequence $\widehat{\boldsymbol{s}}_i = [\alpha_1, \dots, \alpha_{t_i}]^\top$ (see Section 3.1), corresponding to the first t_i stepsizes in $\widehat{\boldsymbol{s}}$. .
- o_t: the integer such that $\sum_{j=1}^{o_t-1} k_j 2^j < t \le \sum_{j=1}^{o_t} k_j 2^j$. Clearly, the length of the $(i+1)$-th subsequence (including the (t_i+1)-th step) is $2^{o_{t_i+1}}$, and $t_{i+1} = t_i + 2^{o_{t_i+1}}$.
- A_t and C_t: $A_t = \sum_{i=1}^{t-1} \alpha_i$ and $C_t = \frac{A_t(A_t+1)}{2}$, where we suppress the dependency on $\widehat{\boldsymbol{s}}$ for notational convenience.
- $\alpha_j(\overline{s}_i)$: the j-th stepsize in the sequence $\overline{\boldsymbol{s}}_i$.

It is also worth noting that Lemma 7 gives

$$2^{o_{t_i+1}} \le 2 \cdot 2^{o_{t_i}} \le 4 t_i^{\frac{1}{c+1}}. \tag{30}$$

Consider any $i \ge 1$. In view of Lemma 4, we know that $\widehat{\boldsymbol{s}}_i$ is primitive. Given that $\{\boldsymbol{x}_j\}_{j=1}^{t_i+1}$ is the GD trajectory with stepsize schedule $\widehat{\boldsymbol{s}}_i$, we see from Definition 2 of the primitive stepsize schedule that

$$A_{t_i+1}(f_{t_i+1} - f^*) + C_{t_i+1}\|\boldsymbol{g}_{t_i+1}\|^2 + \frac{1}{2}\|\boldsymbol{x}_{t_i+1} - \boldsymbol{x}^*\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2,$$

which immediately implies that

$$f_{t_i+1} - f^* \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}}; \tag{31a}$$

$$\|\boldsymbol{g}_{t_i+1}\|^2 \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{2C_{t_i+1}} \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}^2}. \tag{31b}$$

Additionally, by construction we have $\alpha_{t_i+1} = \varphi(x, y)$ due to the concatenation operation, where

$$\begin{aligned}
x &= \sum_{j=1}^{t_i} \alpha_j \ge \frac{1}{36} t_i^{\frac{c+\log_2 \rho}{c+1}} \qquad\qquad ; \\
y &= \sum_{j=t_i+2}^{t_{i+1}} \alpha_j = \sum_{j=1}^{2^{o_{t_i+1}}-1} \alpha_j(\overline{\boldsymbol{s}}_{o_{t_i+1}}) = \rho^{o_{t_i+1}} - 1 \le 2 \cdot t_i^{\frac{\log_2 \rho}{c+1}}.
\end{aligned}$$

Here, both of the inequalities above arise from Lemma 7. It is also easy to observe that $x \ge y$. It then follows that

$$\begin{aligned}
\alpha_{t_i+1} = \varphi(x, y) &= \frac{-(x+y) + \sqrt{(x+y+2)^2 + 4(x+1)(y+1)}}{2} \\
&= \frac{4(xy + 2x + 2y + 2)}{2(x + y + \sqrt{(x+y+2)^2 + 4(x+1)(y+1)})} \\
&\le \frac{xy + 2x + 2y + 2}{x + y + 2} \\
&\le y + 2 \\
&= \rho^{o_{t_i+1}} + 1.
\end{aligned}$$

Moreover, recognizing that

$$\frac{\partial \varphi(x, y)}{\partial x} = \frac{1}{2}\left(-1 + \frac{x + 3y + 4}{\sqrt{x^2 + (6y+8)x + 8y + 8}}\right) \ge 0$$

for all $(x, y) \geq 0$, we immediately obtain

$$\alpha_{t_i+1} = \varphi(x, y) \geq \varphi(y, y) = (\sqrt{2} - 1)y + \sqrt{2}.$$

Invoking (31b) and Lemma 10 over the $(i+1)$-th sub-sequence with $\alpha = \alpha_{t_i+1} \geq (\sqrt{2}-1)y + \sqrt{2}$, we can show that: for any ℓ obeying $t_i + 1 < \ell \leq t_{i+1}$,

$$\begin{aligned}
f_\ell - f_{t_i+1} &\overset{\text{(i)}}{\leq} \frac{3}{2} \cdot (\ell - t_i - 1) \cdot O\left(\max_{t_i+1 \leq j \leq \ell} \alpha_j^2 \|\boldsymbol{g}_j\|^2 \right) \\
&\overset{\text{(ii)}}{\leq} 2^{o_{t_i+1}} \cdot O\left(16^{o_{t_i+1}} \alpha_{t_i+1}^2 \|\boldsymbol{g}_{t_i+1}\|^2\right) \\
&\overset{\text{(iii)}}{\leq} O\left(32^{o_{t_i+1}} \frac{\alpha_{t_i+1}^2}{A_{t_i+1}^2} \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2\right) \\
&\overset{\text{(iv)}}{\leq} O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \cdot t_i^{\frac{5+2\log_2\rho}{c+1}}}{t_i^{\frac{2(c+\log_2\rho)}{c+1}}} \right) \\
&= O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{2(c+\log_2\rho)-5-2\log_2\rho}{c+1}}} \right) \\
&\leq O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{c+\log_2\rho}{c+1}}} \right).
\end{aligned}$$

Here, (i) is valid since, according to (8f) and the fact that $\alpha_j \geq 1$ (as $\varphi(x', y') \geq 1$ for any $x', y' > 0$),

$$f_\ell - f_{t_i+1} = \sum_{j=t_i+1}^{\ell-1} \left(f_{j+1} - f_j\right) \leq \sum_{j=t_i+1}^{\ell-1} \frac{\alpha_j^2 + 2\alpha_j}{2} \|\boldsymbol{g}_j\|^2 \leq \sum_{j=t_i+1}^{\ell-1} \frac{3\alpha_j^2}{2} \|\boldsymbol{g}_j\|^2;$$

(ii) arises from Lemma 10; (iii) is a consequence of (31b); and (iv) invokes Lemma 7, inequality (30), as well as the property that $\alpha_{t_i+1} \leq \rho^{o_{t_i+1}} + 3 = O(\rho^{o_{t_i+1}})$. This taken together with (31a) further results in

$$\begin{aligned}
f_\ell - f^* &= f_\ell - f_{t_i+1} + \left(f_{t_i+1} - f^*\right) \\
&\leq O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{c+\log_2\rho}{c+1}}} + \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}} \right) = O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{c+\log_2\rho}{c+1}}} \right).
\end{aligned}$$

Consequently, we have shown that, for any $\ell \in \cup_{i\geq 1}(t_i, t_{i+1}] = [3, \infty)$,

$$f_\ell - f^* \leq O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}} + \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{c+\log_2\rho}{c+1}}} \right) = O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{c+\log_2\rho}{c+1}}} \right).$$

When $\ell < 3$, it is easily seen that

$$\begin{aligned}
&f_1 - f^* \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{2} \qquad \text{and} \\
&f_2 - f^* \leq f_1 - f^* + \frac{\alpha_1^2 + 2\alpha_1}{2} \|\boldsymbol{g}_1\|^2 \leq (1 + \alpha_1^2 + 2\alpha_1)(f_1 - f^*) \leq \frac{9\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{2},
\end{aligned}$$

where we have made use of (8f). We have thus completed the proof by recalling that $c = 7$.

4 Extension to smooth and strongly convex problems

In this section, we further extend our result to accommodate smooth and strongly convex optimization; that is, we assume that the objective function f in (1) is 1-smooth and μ-strongly convex for some strong convexity parameter $\mu \in (0, 1]$. Here and throughout, we denote by $\kappa = 1/\mu$ the condition number. Our result, which guarantees acceleration of standard GD theory (i.e., $\exp(-\Theta(T/\kappa))$ in an anytime manner, is stated as follows.

Theorem 11. *There is a stepsize schedule $\{\alpha_t\}_{t=1}^{\infty}$, generated without knowing the stopping time, such that the gradient descent iterates* (2) *obey*

$$f(\boldsymbol{x}_T) - f^* \leq O\left(\exp\left(-\frac{CT}{\kappa^{\varsigma}}\right)\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2\right), \tag{32}$$

where $\varsigma = 1/\vartheta = \frac{8}{7+\log_2\rho} < 0.97$, and $C > 0$ is some numerical constant. Here, T denotes an arbitrary stopping time that is unknown a priori.

Proof of Theorem 11. Recall our construction of $\widehat{\boldsymbol{s}}$ in the proof of Theorem 1 (see Section 3.1). According to Theorem 1, there exists a universal constant $C_0 > 0$ such that running GD with the stepsize schedule $\widehat{\boldsymbol{s}}$ achieves

$$f(\boldsymbol{x}_t) - f^* \leq \frac{C_0\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{t^{\vartheta}}.$$

Let us begin by constructing a stepsize schedule tailored to the μ-strongly convex problem. Take $\tau = \tau(\mu)$ to be the smallest integer such that $A_{\tau+1}(\widehat{\boldsymbol{s}}) \geq \frac{4C_0}{\mu} = 4C_0\kappa$. Lemma 7 tells us that

$$\frac{1}{36}\tau^{\vartheta} = \frac{1}{36}\tau^{\frac{c+\log_2\rho}{c+1}} \leq A_{\tau}(\widehat{\boldsymbol{s}}) \leq 4C_0\kappa, \tag{33}$$

which implies that

$$\tau \leq 144C_0\kappa^{\frac{1}{\vartheta}} = 144C_0\kappa^{\varsigma}.$$

Now, let $\widetilde{\boldsymbol{s}} = \boldsymbol{\alpha}_{1:\tau}(\widehat{\boldsymbol{s}})$ (i.e., the first τ stepsizes in $\widehat{\boldsymbol{s}}$), and set $\widetilde{\boldsymbol{s}}^*$ to be the infinite stepsize schedule $[\widetilde{\boldsymbol{s}}^\top, \widetilde{\boldsymbol{s}}^\top, \dots]^\top$; that is, $\alpha_{i\tau+j}(\widetilde{\boldsymbol{s}}^*) = \alpha_j(\widetilde{\boldsymbol{s}}) = \alpha_j(\widehat{\boldsymbol{s}})$ for any $i \geq 0$ and $1 \leq j \leq \tau$.

Next, we would like to show that the claimed result (32) holds with the stepsize schedule $\widehat{\boldsymbol{s}}^*$. In view of Theorem 1, we know that

$$f_j - f^* \leq \frac{C_0\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{j^{\vartheta}} \leq 55C_0\exp\left(-\frac{j}{36C_0\kappa^{\varsigma}}\right)\cdot\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \qquad \text{for all } 1 \leq j \leq \tau; \tag{34}$$

$$f_{\tau+1} - f^* \leq \frac{C_0\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\tau^{\theta}} = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{144\kappa^{\varsigma\cdot\vartheta}} = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{144\kappa} = \frac{\mu\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{144}, \tag{35}$$

where we have invoked (33). Observing that $f_{\tau+1} - f^* \geq \frac{\mu\|\boldsymbol{x}_{\tau+1} - \boldsymbol{x}^*\|^2}{2}$ due to μ-strong convexity, we have

$$\|\boldsymbol{x}_{\tau+1} - \boldsymbol{x}^*\|^2 \leq \frac{1}{72}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2.$$

Invoking similar arguments reveals that: for any $i \geq 1$ and $1 \leq j \leq \tau$, one has

$$\frac{\mu}{2}\|\boldsymbol{x}_{i\tau+1} - \boldsymbol{x}^*\|^2 \leq f_{i\tau+1} - f^* \leq \frac{\mu\|\boldsymbol{x}_{(i-1)\tau+1} - \boldsymbol{x}^*\|^2}{4}$$

$$\text{and} \qquad f_{i\tau+j} - f^* \leq \frac{C_0\|\boldsymbol{x}_{i\tau+1} - \boldsymbol{x}^*\|^2}{j^{\vartheta}} \leq C_0\|\boldsymbol{x}_{i\tau+1} - \boldsymbol{x}^*\|^2.$$

As a result, we can deduce that

$$\begin{aligned}
\|\boldsymbol{x}_{i\tau+1} - \boldsymbol{x}^*\|^2 &\leq \frac{1}{2}\|\boldsymbol{x}_{(i-1)\tau+1} - \boldsymbol{x}^*\|^2 \\
&\leq \left(\frac{1}{2}\right)^i\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \\
&\leq \exp\left(-\log 2\cdot\frac{i\tau+1}{2\tau}\right)\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \\
&\leq \exp\left(-\frac{i\tau+1}{576C_0\kappa^{\varsigma}}\right)\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2,
\end{aligned}$$

and as a result,

$$f_{i\tau+j} - f^* \le C_0 \|\boldsymbol{x}_{i\tau+1} - \boldsymbol{x}^*\|^2 \le C_0 \exp\left(-\frac{i\tau+1}{576 C_0 \kappa^{\varsigma}}\right) \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \le C_0 \exp\left(-\frac{i\tau+j}{1152 C_0 \kappa^{\varsigma}}\right) \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2. \tag{36}$$

To finish up, combine (34) and (36) to arrive at

$$f_T - f^* \le 55 C_0 \exp\left(-\frac{CT}{\kappa^{\varsigma}}\right) \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \qquad \text{for any } T \ge 1,$$

where $C = \frac{1}{1152 C_0}$. This concludes the proof. □

Acknowledgements

YC is supported in part by the Alfred P. Sloan Research Fellowship, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grant CCF-2221009. SSD is supported in part by the Alfred P. Sloan Research Fellowship, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, and NSF IIS 2229881. JDL acknowledges support of NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994. We would like to thank Ernest Ryu for helpful discussions and references regarding the literature.

References

Altschuler, J. (2018). *Greed, hedging, and acceleration in convex optimization.* PhD thesis, Massachusetts Institute of Technology.

Altschuler, J. M. and Parrilo, P. A. (2023a). Acceleration by stepsize hedging i: Multi-step descent and the silver stepsize schedule. *arXiv preprint arXiv:2309.07879.*

Altschuler, J. M. and Parrilo, P. A. (2023b). Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization. *arXiv preprint arXiv:2309.16530.*

Beck, A. (2017). *First-order methods in optimization.* SIAM.

Daccache, A., Glineur, F., and Hendrickx, J. (2019). *Performance estimation of the gradient method with fixed arbitrary step sizes.* PhD thesis, Master's thesis, Université Catholique de Louvain.

Das Gupta, S., Van Parys, B. P., and Ryu, E. K. (2024). Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods. *Mathematical Programming*, 204(1):567–639.

Drori, Y. and Teboulle, M. (2014). Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482.

Eloi, D. and Glineur, F. (2022). *Worst-case functions for the gradient method with fixed variable step sizes.* PhD thesis, Master's thesis, Université Catholique de Louvain.

Grimmer, B. (2024). Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608.

Grimmer, B., Shu, K., and Wang, A. L. (2023). Accelerated gradient descent via long steps. *arXiv preprint arXiv:2309.09961.*

Grimmer, B., Shu, K., and Wang, A. L. (2024a). Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *arXiv preprint arXiv:2403.14045.*

Grimmer, B., Shu, K., and Wang, A. L. (2024b). Composing optimized stepsize schedules for gradient descent. *arXiv preprint arXiv:2410.16249*.

Kornowski, G. and Shamir, O. (2024). Open problem: Anytime convergence rate of gradient descent. In *Conference on Learning Theory*, volume 247, pages 5335–5339.

Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

Rotaru, T., Glineur, F., and Patrinos, P. (2024). Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions. *arXiv preprint arXiv:2406.17506*.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2017). Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345.

Teboulle, M. and Vaisbourd, Y. (2023). An elementary approach to tight worst case complexity analysis of gradient based methods. *Mathematical Programming*, 201(1):63–96.

Zhang, Z. and Jiang, R. (2024). Accelerated gradient descent by concatenation of stepsize schedules. *arXiv preprint arXiv:2410.12395*.

A Proof of Lemma 7

First, let us look at the case with $o_t = 1$, for which we have $t \le 2k_1 = 2^{c+1}$. Given that $\varphi(x,y) > 1$ for all $x, y \ge 0$, we can easily verify that

$$A_{t+1}(\widehat{\boldsymbol{s}}) \ge t \ge \frac{1}{36} t^{\frac{c+\log_2 \rho}{c+1}}.$$

It is also easily seen that $2^{o_t} = 2 \le 2t^{\frac{1}{c+1}}$.

Now, let us turn to the case where $o_t \ge 2$. Let $m \in [1, k_{o_t}]$ be the integer such that $\sum_{j=1}^{o_t-1} k_j 2^j + (m-1)\cdot 2^{o_t} < t \le \sum_{j=1}^{o_t-1} k_j 2^j + m \cdot 2^{o_t}$. By definition, we have

$$t \le \sum_{j=1}^{o_t-1} k_j 2^j + m 2^{o_t} \le 2 \cdot 2^{(c+1)(o_t-1)} + m 2^{o_t};$$

$$A_{t+1}(\widehat{\boldsymbol{s}}) \ge \sum_{j=1}^{o_t-1} (\rho^j - 1)\cdot k_j + (m-1)(\rho^{o_t} - 1) \ge \frac{1}{2} \cdot 2^{(c+\log_2 \rho)(o_t-1)} + \frac{m-1}{2}\rho^{o_t},$$

where the second line invokes Lemma 6.

- If $m 2^{o_t} \le 2^{(c+1)(o_t-1)}$, then we have $t \le 3 \cdot 2^{(c+1)(o_t-1)}$, which means that $A_{t+1}(\widehat{\boldsymbol{s}}) \ge \frac{1}{2} \cdot 2^{(c+\log_2 \rho)(o_t-1)} \ge \frac{1}{18} t^{\frac{c+\log_2 \rho}{c+1}}$.
- If $m 2^{o_t} > 2^{(c+1)(o_t-1)}$ — i.e, $2^{o_t c} \ge m > 2^{o_t c - c - 1} \ge 1$ — then one has

$$t^{\frac{c+\log_2 \rho}{c+1}} < 9(m 2^{o_t})^{\frac{c+\log_2 \rho}{c+1}} \le 9 \cdot m \rho^{o_t} \le 36 \cdot \frac{m-1}{2}\rho^{o_t} \le 36 A_{t+1}(\widehat{\boldsymbol{s}}).$$

Putting these two cases together establishes the claim (16).

Regarding the second claim, in the case where $o_t \ge 2$, we have

$$t \ge \sum_{j=1}^{o_t-1} k_j 2^j = \sum_{j=1}^{o_t-1} 2^{(c+1)j} \ge 2^{(c+1)(o_t-1)}, \tag{37}$$

thus indicating that $2t^{\frac{1}{c+1}} \ge 2^{o_t}$.

B Proof of preliminary facts from Zhang and Jiang (2024)

B.1 Proof of Lemma 3

As mentioned previously, this lemma was established by Zhang and Jiang (2024). We present the proof for completeness.

To begin with, we single out the following lemma, originally established by Zhang and Jiang (2024, Lemma 3.1), that plays a key role in the proof of Lemma 3. We shall provide a proof in Appendix B.2.

Lemma 12. *(Zhang and Jiang (2024, Lemma 3.1)) Assume that $\boldsymbol{\alpha}_{1:\ell-1}$ is primitive. For any $\alpha \in [1, A_\ell + 2)$, if we set $\alpha_0 = \alpha$, then it holds that*

$$f_0 - f_\ell \ge \frac{A_\ell + 3\alpha - 2\alpha^2}{2(A_\ell + 2 - \alpha)}\|\boldsymbol{g}_0\|^2 + \frac{2A_\ell^2 + 3A_\ell + \alpha}{2(A_\ell + 2 - \alpha)}\|\boldsymbol{g}_\ell\|^2.$$

Next, in view of the definition of the primitive stepsize schedule (cf. Definition 2), we can easily see that

$$x(f_\ell - f^*) + \frac{x(x+1)}{2}\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{\ell-1} \alpha_i \left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right),$$

$$y(f_k - f^*) + \frac{y(y+1)}{2}\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \le \frac{1}{2}\|\boldsymbol{x}_{\ell+1} - \boldsymbol{x}^*\|^2 + \sum_{i=\ell+1}^{k-1} \alpha_i \left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right),$$

where we take $x = A_\ell$ and $y = A_k - A_{\ell+1}$ for notational simplicity. Given that

$$z \coloneqq x + y + \alpha = A_\ell + (A_k - A_{\ell+1}) + \alpha_\ell = A_k,$$

Lemma 12 tells us that

$$(x+\alpha)(f_k - f_\ell) \le -\frac{(x+\alpha)(y+3\alpha-2\alpha^2)}{2(y+2-\alpha)}\|\boldsymbol{g}_\ell\|^2 - \frac{(x+\alpha)(2y^2+3y+\alpha)}{2(y+2-\alpha)}\|\boldsymbol{g}_k\|^2. \tag{38}$$

Adding the above three inequalities and utilizing $z = x + y + \alpha$ yield

$$L_1 + L_2 + L_3 + L_4 \le R_1 + R_2 + R_3 + R_4, \tag{39}$$

where

$$\begin{aligned}
L_1 &= z(f_k - f^*) + \frac{z(z+1)}{2}\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 = A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2;\\
L_2 &= -\alpha(f_\ell - f^*) + \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2;\\
L_3 &= \frac{x(x+1)}{2}\|\boldsymbol{g}_\ell\|^2;\\
L_4 &= \frac{y(y+1) - z(z+1)}{2}\|\boldsymbol{g}_k\|^2;\\
R_1 &= \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right);\\
R_2 &= \frac{1}{2}\|\boldsymbol{x}_{\ell+1} - \boldsymbol{x}^*\|^2 - \alpha\big(f_\ell - f^* - \langle \boldsymbol{g}_\ell, \boldsymbol{x}_\ell - \boldsymbol{x}^* \rangle\big) - \frac{1}{2}\alpha^2\|\boldsymbol{g}_\ell\|^2;\\
R_3 &= \left(-\frac{\alpha}{2} + \frac{\alpha^2}{2} - \frac{(x+\alpha)(y+3\alpha-2\alpha^2)}{(y+2-\alpha)} \right)\|\boldsymbol{g}_\ell\|^2;\\
R_4 &= -\frac{(x+\alpha)(2y^2+3y+\alpha)}{y+2-\alpha}\|\boldsymbol{g}_k\|^2.
\end{aligned}$$

We now proceed to simplify (39). Firstly, it is readily seen that

$$\begin{aligned}
L_2 - R_2 &= \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2 - \frac{1}{2}\|\boldsymbol{x}_{\ell+1} - \boldsymbol{x}^*\|^2 - \alpha\langle \boldsymbol{g}_\ell, \boldsymbol{x}_\ell - \boldsymbol{x}^* \rangle + \frac{1}{2}\alpha^2\|\boldsymbol{g}_\ell\|^2\\
&= \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2 - \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^* - \alpha\boldsymbol{g}_\ell\|^2 - \alpha\langle \boldsymbol{g}_\ell, \boldsymbol{x}_\ell - \boldsymbol{x}^* \rangle + \frac{1}{2}\alpha^2\|\boldsymbol{g}_\ell\|^2\\
&= 0.
\end{aligned}$$

Secondly, recalling our specific choice $\alpha = \varphi(x,y) = \frac{-(x+y)+\sqrt{(x+y+2)^2+4(x+1)(y+1)}}{2}$, we can easily verify that

$$\alpha^2 + (x+y)\alpha - (xy + 2x + 2y + 2) = 0.$$

This allows one to demonstrate that

$$\begin{aligned}
L_3 - R_3 &= \left(\frac{x(x+1)}{2} + \frac{\alpha}{2} - \frac{\alpha^2}{2} + \frac{(x+\alpha)(y+3\alpha-2\alpha^2)}{2(y+2-\alpha)} \right)\|\boldsymbol{g}_\ell\|^2 = 0;\\
L_4 - R_4 &= \left(\frac{y(y+1) - z(z+1)}{2} + \frac{(x+\alpha)(2y^2+3y+\alpha)}{y+2-\alpha} \right)\|\boldsymbol{g}_k\|^2 = 0.
\end{aligned}$$

Substitution into (39) then results in $L_1 \le R_1$, namely,

$$
\begin{aligned}
A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 &\le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right), \\
&\le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2, \qquad (40)
\end{aligned}
$$

where the last inequality comes from (8a). This completes the proof.

B.2 Proof of Lemma 12

Once again, this lemma has been proven in Zhang and Jiang (2024, Lemma 3.1), and we present the proof for completeness.

According to the definition of the primitive stepsize schedule, we have

$$
A_\ell(f_\ell - f^*) + C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{\ell-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle) + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right). \quad (41)
$$

Recall from the basic properties (8) that

$$
\begin{aligned}
f_i - f_\ell &\le \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_\ell\rangle - \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_\ell\|^2, \\
f_i - f_0 &\le \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_0\rangle - \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_0\|^2,
\end{aligned}
$$

which allow us to derive

$$
\begin{aligned}
&f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \\
&\quad\le f_\ell - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_\ell - \boldsymbol{x}_i + \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 - \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_\ell\|^2 \\
&\quad= f_\ell - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_\ell - \boldsymbol{x}^*\rangle + \langle \boldsymbol{g}_i, \boldsymbol{g}_\ell\rangle - \frac{1}{2}\|\boldsymbol{g}_\ell\|^2,
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
&f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \\
&\quad\le f_0 - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^*\rangle + \langle \boldsymbol{g}_i, \boldsymbol{g}_0\rangle - \frac{1}{2}\|\boldsymbol{g}_0\|^2.
\end{aligned}
$$

As a result, we can take advantage of these properties to deduce that

$$
\begin{aligned}
&\sum_{i=1}^{\ell-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right) \\
&\quad\le A_\ell(f_\ell - f^*) - \sum_{i=1}^{\ell}\alpha_i\langle \boldsymbol{g}_i, \boldsymbol{x}_\ell - \boldsymbol{x}^*\rangle + \sum_{i=1}^{\ell-1}\alpha_i\langle \boldsymbol{g}_i, \boldsymbol{g}_\ell\rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_\ell\|^2 \\
&\quad= A_\ell(f_\ell - f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_\ell - \boldsymbol{x}^*\rangle + \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{g}_\ell\rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_\ell\|^2, \qquad (42a)
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
&\sum_{i=1}^{\ell-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right) \\
&\quad\le A_\ell(f_0 - f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_0 - \boldsymbol{x}^*\rangle + \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{g}_0\rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_0\|^2. \qquad (42b)
\end{aligned}
$$

Combine (42a) and (42b) to arrive at

$$
\begin{aligned}
&\sum_{i=1}^{\ell-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right)\\
&\le \frac{1}{2}\left(A_\ell(f_0 + f_\ell - 2f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_0 + \boldsymbol{x}_\ell - 2\boldsymbol{x}^*\rangle + \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_\ell\|^2 - \frac{A_\ell}{2}\|\boldsymbol{g}_0\|^2\right)\\
&= \frac{A_\ell}{2}\left(f_0 + f_\ell - 2f^* - \frac{\|\boldsymbol{g}_\ell\|^2}{2} - \frac{\|\boldsymbol{g}_0\|^2}{2}\right)\\
&\qquad - \frac{1}{2}\langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_1 + \alpha\boldsymbol{g}_0 + \boldsymbol{x}_\ell - 2\boldsymbol{x}^*\rangle + \frac{1}{2}\langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle - \frac{1}{2}\alpha\big(\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \|\boldsymbol{g}_0\|^2\big)\\
&= \frac{A_\ell}{2}\left(f_0 + f_\ell - 2f^* - \frac{\|\boldsymbol{g}_\ell\|^2}{2} - \frac{\|\boldsymbol{g}_0\|^2}{2}\right)\\
&\qquad - \frac{1}{2}\langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_1 + \boldsymbol{x}_\ell - 2\boldsymbol{x}^*\rangle - \frac{1}{2}\alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_1 - \boldsymbol{x}_\ell\rangle + \frac{1}{2}\langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle - \frac{1}{2}\alpha\big(\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \|\boldsymbol{g}_0\|^2\big)\\
&= \frac{A_\ell}{2}\left(f_0 + f_\ell - 2f^* - \frac{\|\boldsymbol{g}_\ell\|^2}{2} - \frac{\|\boldsymbol{g}_0\|^2}{2}\right)\\
&\qquad - \frac{1}{2}\big(\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2\big) - \frac{1}{2}\alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_1 - \boldsymbol{x}_\ell\rangle + \frac{1}{2}\langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle - \frac{1}{2}\alpha\big(\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \|\boldsymbol{g}_0\|^2\big).
\end{aligned}
$$

Adding this inequality and (41), we further reach

$$
\begin{aligned}
&A_\ell(f_\ell - f^*) + C_\ell\|\boldsymbol{g}_\ell\|^2\\
&\le \frac{1}{2}A_\ell\left(f_0 + f_\ell - 2f^* - \frac{\|\boldsymbol{g}_\ell\|^2}{2} - \frac{\|\boldsymbol{g}_0\|^2}{2}\right) - \frac{1}{2}\alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_1 - \boldsymbol{x}_\ell\rangle + \frac{1}{2}\langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle - \frac{1}{2}\alpha\big(\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \|\boldsymbol{g}_0\|^2\big).
\end{aligned}
$$

Rearrange terms to arrive at

$$
\begin{aligned}
&A_\ell(f_0 - f_\ell)\\
&\ge 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_1 - \boldsymbol{x}_\ell\rangle - \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2\\
&= 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \alpha\boldsymbol{g}_0 - \boldsymbol{x}_\ell\rangle - \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2\\
&= 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_\ell\rangle - \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2 - \alpha^2\|\boldsymbol{g}_0\|^2\\
&= 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, (\alpha-1)\boldsymbol{g}_0 - \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2 - \alpha^2\|\boldsymbol{g}_0\|^2. \qquad (43)
\end{aligned}
$$

The next step is to bound the term $\langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, (\alpha-1)\boldsymbol{g}_0 - \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle$. Towards this, we recall from (8) that

$$
\begin{aligned}
(\alpha-1)(f_0 - f_\ell) &\le (\alpha-1)\langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_\ell\rangle - \frac{\alpha-1}{2}\|\boldsymbol{g}_0 - \boldsymbol{g}_\ell\|^2;\\
f_\ell - f_0 &\le -\langle \boldsymbol{g}_\ell, \boldsymbol{x}_0 - \boldsymbol{x}_\ell\rangle - \frac{1}{2}\|\boldsymbol{g}_0 - \boldsymbol{g}_\ell\|^2.
\end{aligned}
$$

Adding the preceding two inequalities gives

$$
\begin{aligned}
(\alpha-2)(f_0 - f_\ell) &\le \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, (\alpha-1)\boldsymbol{g}_0 - \boldsymbol{g}_\ell\rangle - \frac{\alpha}{2}\|\boldsymbol{g}_0 - \boldsymbol{g}_\ell\|^2\\
&= \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, (\alpha-1)\boldsymbol{g}_0 - \boldsymbol{g}_\ell\rangle - \frac{\alpha}{2}(\|\boldsymbol{g}_0\|^2 + \|\boldsymbol{g}_\ell\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle,
\end{aligned}
$$

thus indicating that

$$
\langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, (\alpha-1)\boldsymbol{g}_0 - \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle \ge (\alpha-2)(f_0 - f_\ell) + \frac{\alpha}{2}(\|\boldsymbol{g}_0\|^2 + \|\boldsymbol{g}_\ell\|^2). \qquad (44)
$$

Substitution into (43) then leads to

$$A_\ell(f_0 - f_\ell) \geq 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + (\alpha - 2)(f_0 - f_\ell) + \frac{\alpha}{2}(\|\boldsymbol{g}_0\|^2 + \|\boldsymbol{g}_\ell\|^2) + \alpha\|\boldsymbol{g}_0\|^2 - \alpha^2\|\boldsymbol{g}_0\|^2.$$

Rearranging terms and using $C_\ell = \frac{A_\ell(A_\ell+1)}{2}$, we are left with

$$(A_\ell + 2 - \alpha)(f_0 - f_\ell) \geq \left(A_\ell^2 + \frac{3A_\ell}{2} + \frac{\alpha}{2}\right)\|\boldsymbol{g}_\ell\|^2 + \left(\frac{A_\ell}{2} + \frac{\alpha}{2} + \alpha - \alpha^2\right)\|\boldsymbol{g}_0\|^2. \tag{45}$$

Dividing both sides of the above display by $(A_\ell + 2 - \alpha)$, we conclude the proof.