Proceedings of Machine Learning Research vol ???:1–23, 2025 38th Annual Conference on Learning Theory

# Anytime Acceleration of Gradient Descent

**Zihan Zhang** ZIHANZ46@UW.EDU
*University of Washington, Seattle, WA 98195, USA.*

**Jason D. Lee** JASONLEE@PRINCETON.EDU
*Princeton University, Princeton, NJ 08544, USA.*

**Simon S. Du** SSDU@CS.WASHINGTON.EDU
*University of Washington, Seattle, WA 98195, USA.*

**Yuxin Chen** YUXINC@WHARTON.UPENN.EDU
*University of Pennsylvania, Philadelphia, PA 19104, USA.*

**Editors:** Nika Haghtalab and Ankur Moitra

## 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 $O\big(T^{-\frac{2\log_2\rho}{1+\log_2\rho}}\big) \approx O(T^{-1.119})$ for any stopping time $T$, where $\rho = \sqrt{2}+1$ is the silver ratio and 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 $\exp(-\Omega(T/\kappa^{0.893}))$ for smooth and strongly convex optimization, with $\kappa$ being the condition number.

## 1. Introduction

Consider the standard problem of smooth convex optimization:

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

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, 2018). Somewhat surprisingly, a recent strand of work (Teboulle and Vaisbourd, 2023; Altschuler and Parrilo, 2023b,a; Altschuler, 2018; Grimmer, 2024; Grimmer et al., 2023; Rotaru et al., 2024; Grimmer et al., 2024a) 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

$$\rho := 1 + \sqrt{2}$$

is the silver ratio 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]

$$f(\boldsymbol{x}_T) - f^* \leq O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{T^{\vartheta}}\right) \qquad \text{with } \vartheta = \frac{2\log_2 \rho}{1 + \log_2 \rho} \approx 1.119 \tag{4}$$

*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)). A key ingredient underlying our algorithm design is to ensure that the sizes of the gradients in intermediate iterations are well-controlled, so that the intermediate steps do not overshoot. From a technical perspective, our result suggests that the application of interpolation inequalities might offer a promising lens to sharpen the analysis for standard convex optimization algorithms.

1. Throughout this paper, we use $\|\cdot\|$ to denote the $\ell_2$ norm.

**Remark 2** *Comparing the $O(T^{-\frac{2\log_2(\rho)}{1+\log_2(\rho)}})$ rate in Theorem 1 with the $O(T^{-\log_2(\rho)})$ rate in Altschuler and Parrilo (2023a), one might naturally ask two questions: (i) Can the* anytime *convergence rates be improved to match the ones derived for a fixed given $T$? (ii) What are the optimal convergence rates for both the* anytime *setting and the fixed $T$ setting? To address these two questions, a key challenge lies in identifying an appropriate class of convex and smooth functions for constructing tight lower bounds, particularly in high-dimensional settings. We leave these for future studies.*

**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 is described in Section 3, while the proof of Theorem 1 is provided in Section 4. In Section 5, building upon the standard approach from Roulet and d'Aspremont (2020), 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

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

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 $[\alpha_\ell, \dots, \alpha_k]^\top$, 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 $f(\cdot)$:

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

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

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

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

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

$$\begin{aligned} &f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big) - f(\boldsymbol{x}) \\ &\quad \le \alpha \big\langle \nabla f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big), \nabla f(\boldsymbol{x}) \big\rangle - \frac{1}{2}\big\|\nabla f(\boldsymbol{x}) - \nabla f\big(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})\big)\big\|^2. \end{aligned} \tag{8e}$$

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

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

**Primitive stepsize schedule and concatenation.** Our algorithm is built upon the notion of "primitive stepsize schedule" as introduced in Zhang and Jiang (2024, Definition 3), detailed below.

**Definition 3 (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 (see (6) for the definition of $A_k$ and $C_k$)*

$$\begin{aligned} &A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\ &\quad \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right) \qquad (9a) \\ &\quad \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2. \qquad (9b) \end{aligned}$$

When $k = 1$, (9a) holds trivially, which means that the null sequence is a primitive stepsize schedule. At a high level, these definitions are motivated by merging short stepsize sequences into a longer stepsize sequence while preserving convergence guarantees. The following lemma, derived by Zhang and Jiang (2024), makes precise this key property; for completeness, we provide a proof in Appendix E.1.

**Lemma 4** *(Zhang and Jiang (2024, Theorem 3.1)) Consider a stepsize schedule $\{\alpha_t\}_{t\ge1}$. 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*

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

*Then,* $\boldsymbol{\alpha}_{1:k-1} = [\alpha_1, \dots, \alpha_{k-1}]$ *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 4 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:

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

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

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

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

**Lemma 5** *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 4 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 $\widehat{\boldsymbol{s}}$. ■

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

**Definition 6 (Silver stepsize schedule)** *Let* $\overline{\boldsymbol{s}}_0 = [\,]$ *be the null sequence. And for each* $i \geq 1$*, set* $\overline{\boldsymbol{s}}_i = \mathsf{concat}(\overline{\boldsymbol{s}}_{i-1}, \overline{\boldsymbol{s}}_{i-1})$*. Then* $\overline{\boldsymbol{s}}_i$ *is said to be the* $i$*-th order silver stepsize schedule, with the (limiting) silver stepsize schedule given by* $\overline{\boldsymbol{s}} \coloneqq \lim_{i\to\infty} \overline{\boldsymbol{s}}_i$*.*

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 7** *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 4 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. ■

## 3. The proposed stepsize schedule

### 3.1. Construction of our stepsize schedule

Armed with the silver stepsize schedules $\{\overline{\boldsymbol{s}}_j\}_{j\geq 0}$ introduced in Definition 6 — which serve as basic primitive sequences — we can readily present the proposed stepsize schedule.

Choose some positive quantity $c \geq 1$. While we shall keep $c$ as a general quantity throughout most of the proof, it will be taken to be $c = \log_2 \rho$ at the last step of our proof of the main theorems. Also, set

$$k_0 = M_0 = 0, \qquad k_i = \left\lfloor 2 \cdot 2^{ci} \right\rfloor, \qquad \text{and} \qquad M_i = \sum\nolimits_{j=1}^{i} k_i, \qquad i = 1, 2, \dots \tag{15}$$

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 6. 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).

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 8** *For any $t \geq 1$, it holds that*

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

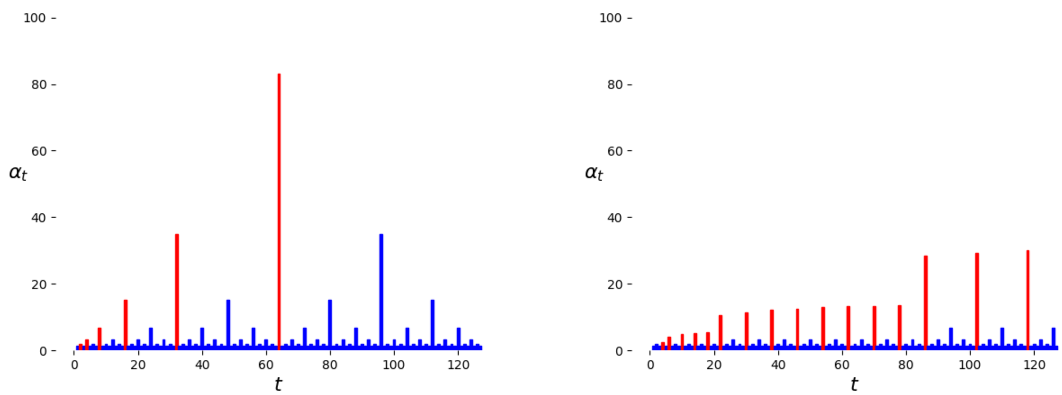
*where $A_t(\widehat{\boldsymbol{s}})$ is defined in (7). Moreover, letting $o_t$ denote the integer obeying $\sum_{j=1}^{o_t - 1} k_j 2^j < t \leq \sum_{j=1}^{o_t} k_j 2^j$, one has*

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

### 3.2. A glimpse of high-level ideas

Before proceeding, let us briefly outline the high-level intuition behind our design and analysis of the stepsize schedule.

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 $\Omega(t^{\frac{\log_2 \rho}{\log_2 \rho+1}})$.

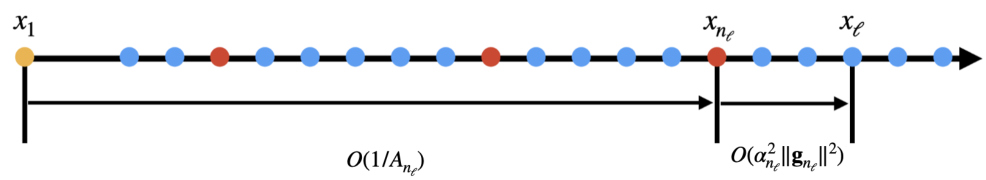



**Convergence at the join steps.** As proven recently by Zhang and Jiang (2024); Grimmer et al. (2024b), certain desirable stepsize schedules with different lengths can be concatenated — with a properly chosen join step — into a longer stepsize schedule while ensuring fast convergence at the last step, 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 step 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, by merging short stepsize schedules into a longer stepsize schedule, we could construct a stepsize schedule with $\Omega(t^{1-\epsilon_1})$ join steps and an aggregate stepsize $\Omega(t^{1+\epsilon_2})$ for proper constants $\epsilon_1, \epsilon_2 > 0$, which enables a convergence rate of $o(t^{-1})$ at each join step.

**Closeness to the join steps.** 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 $\ell$, and denote by $n_\ell$ the iteration number of the closest join step below $\ell$; 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 $n_\ell$, for which the silver stepsize schedule provides effective control of the weighted gradient norm (see Lemma 10 and 11).

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




## 4. Analysis

### 4.1. 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). The proof can be found in Appendix B.

**Lemma 9** *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*

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

Furthermore, the result in Lemma 9 allows us to control the gradient norm at the last step, provided that a primitive stepsize schedule is adopted. The proof is deferred to Appendix C.

**Lemma 10** *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*

$$C_k\|\boldsymbol{g}_k\|^2 \le \left(\frac{\alpha_0^2}{2} + \frac{(A_k+1)^2}{2} - \alpha_0 - \frac{A_k}{2}\right)\|\boldsymbol{g}_0\|^2; \tag{19}$$

$$f_k - f_0 \le \frac{1}{A_k}\left(\frac{1}{2}\alpha_0^2 - \frac{A_k}{2} - \alpha_0 + \frac{1}{2}\right)\|\boldsymbol{g}_0\|^2. \tag{20}$$

Additionally, the following lemma enables effective control of the gradient norms in all intermediate steps. The proof is postponed to Appendix D.

**Lemma 11** *Consider $i \ge 1$ and $\alpha \ge 0$, and let $k = 2^i$. Denote by $\bar{\boldsymbol{s}}_i = [\alpha_1, \ldots, \alpha_{k-1}]^\top$ the $i$-th order silver stepsize schedule. Fix $\boldsymbol{x}_0$, set $\alpha_0 = \alpha$ and let $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha_0\boldsymbol{g}_0$. If $\alpha \ge (\sqrt{2}-1)A_k + \sqrt{2}$, then one has*

$$f_\ell - f_0 \le 432\alpha^2\|\boldsymbol{g}_0\|^2$$

*for any $\ell$ obeying $1 \le \ell \le k-1$.*

### 4.2. 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 $t_{i+1} = t_i + 2^{o_{t_i+1}} \le 3t_i$.

- $A_t$ 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 8 gives

$$2^{o_{t_i+1}} \le 2 \cdot 2^{o_{t_i}} \le 4t_i^{\frac{1}{c+1}}. \tag{21}$$

Consider any $i \ge 1$. In view of Lemma 5, 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 3 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

$$f_{t_i+1} - f^* \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}}; \tag{22a}$$

$$\|\boldsymbol{g}_{t_i+1}\|^2 \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{2C_{t_i+1}} \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}^2}. \tag{22b}$$

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

$$x = \sum_{j=1}^{t_i} \alpha_j \ge \frac{1}{36} t_i^{\frac{c+\log_2 \rho}{c+1}} \qquad ;$$

$$y = \sum_{j=t_i+2}^{t_{i+1}} \alpha_j = \sum_{j=1}^{2^{o_{t_i+1}}-1} \alpha_j(\overline{\boldsymbol{s}}_{o_{t_i+1}}) = \rho^{o_{t_i+1}} - 1 \le 2 \cdot t_i^{\frac{\log_2 \rho}{c+1}}.$$

Here, both of the inequalities above arise from Lemma 8. It is also easy to observe that $x \geq y$. It then follows that

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

Moreover, recognizing that

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

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 Lemma 11 over the $(i+1)$-th sub-sequence with $\alpha = \alpha_{t_i+1} \geq (\sqrt{2}-1)y + \sqrt{2}$, we can show, for any $\ell$ obeying $t_i + 1 < \ell \leq t_{i+1}$, that

$$\begin{aligned}
f_\ell - f_{t_i+1} \leq 432\alpha_{t_i+1}^2 \|\boldsymbol{g}_{t_i+1}\|^2 &\overset{\text{(i)}}{\leq} O\left(\frac{\alpha_{t_i+1}^2}{A_{t_i+1}^2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2\right) \\
&\overset{\text{(ii)}}{\leq} O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 t_i^{\frac{2\log_2\rho}{c+1}}}{t_i^{\frac{2(c+\log_2\rho)}{c+1}}}\right) \\
&= O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{2(c+\log_2\rho)-2\log_2\rho}{c+1}}}\right) \\
&= O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{2c}{c+1}}}\right).
\end{aligned}$$

Here, (i) arises from (22b), whereas (ii) invokes Lemma 8, inequality (21), as well as the property that

$$\alpha_{t_i+1} \leq \rho^{o_{t_i+1}} + 3 = O(\rho^{o_{t_i+1}}) \leq O\big(t_i^{\frac{\log_2\rho}{c+1}}\big).$$

This taken together with (22a) further results in

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

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

$$f_\ell - f^* \leq O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{A_{t_i+1}} + \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{2c}{c+1}}}\right) \leq O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{2c}{c+1}}} + \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{c+\log_2\rho}{c+1}}}\right) \tag{23}$$

$$= O\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\ell^{\frac{2\log_2\rho}{1+\log_2\rho}}}\right), \tag{24}$$

where the last line follows by taking $c = \log_2 \rho$.

It remains to justify the advertised result when $\ell < 3$. Towards this end, it is easily seen that

$$f_1 - f^* \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{2} \qquad \text{and}$$

$$f_2 - f^* \leq f_1 - f^* + \frac{\alpha_1^2 + 2\alpha_1}{2}\|\boldsymbol{g}_1\|^2 \leq (1 + \alpha_1^2 + 2\alpha_1)(f_1 - f^*) \leq \frac{9\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{2},$$

where we have made use of the following fact

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

We have thus completed the proof.

## 5. Extension to smooth and strongly convex problems

Based on the standard approach from Roulet and d'Aspremont (2020), we further extend our result to accommodate smooth and strongly convex optimization. We assume that the objective function $f$ in (1) is 1-smooth and $\mu$-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., $\exp(-\Omega(T/\kappa))$ in an anytime manner, is stated as follows.

**Theorem 12** *There is a stepsize schedule $\{\alpha_t\}_{t=1}^{\infty}$, generated without knowing the stopping time, such that the gradient descent iterates (2) obey*

$$f(\boldsymbol{x}_T) - f^* \leq O\left(\exp\left(-\frac{CT}{\kappa^{\varsigma}}\right)\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2\right), \tag{25}$$

*where $\varsigma = 1/\vartheta = \frac{1+\log_2\rho}{2\log_2\rho} < 0.893$, and $C > 0$ is some numerical constant. Here, $T$ denotes an arbitrary stopping time that is unknown a priori.*

**Proof** 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 $\mu$-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 8 tells us that

$$\frac{1}{36}\tau^{\vartheta} = \frac{1}{36}\tau^{\frac{c+\log_2\rho}{c+1}} \leq A_\tau(\widehat{\boldsymbol{s}}) \leq 4C_0\kappa, \tag{26}$$

which implies that

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

Now, let $\widetilde{\boldsymbol{s}} = \boldsymbol{\alpha}_{1:\tau}(\widehat{\boldsymbol{s}})$ (i.e., the first $\tau$ 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 $i \ge 0$ and $1 \le j \le \tau$.

Next, we would like to show that the claimed result (25) holds with the stepsize schedule $\widehat{\boldsymbol{s}}^*$. In view of Theorem 1, we know that

$$f_j - f^* \le \frac{C_0 \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{j^{\vartheta}} \le 55 C_0 \exp\left(-\frac{j}{36 C_0 \kappa^{\varsigma}}\right) \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \qquad \text{for all } 1 \le j \le \tau; \tag{27}$$

$$f_{\tau+1} - f^* \le \frac{C_0 \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{\tau^{\theta}} = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{144 \kappa^{\varsigma \cdot \vartheta}} = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{144 \kappa} = \frac{\mu \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2}{144}, \tag{28}$$

where we have invoked (26). Observing that $f_{\tau+1} - f^* \ge \frac{\mu \|\boldsymbol{x}_{\tau+1} - \boldsymbol{x}^*\|^2}{2}$ due to $\mu$-strong convexity, we have

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

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

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

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

As a result, we can deduce that

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

and as a result,

$$\begin{aligned}
f_{i\tau+j} - f^* \le C_0 \|\boldsymbol{x}_{i\tau+1} - \boldsymbol{x}^*\|^2 &\le C_0 \exp\left(-\frac{i\tau + 1}{576 C_0 \kappa^{\varsigma}}\right) \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 \\
&\le C_0 \exp\left(-\frac{i\tau + j}{1152 C_0 \kappa^{\varsigma}}\right) \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2.
\end{aligned} \tag{29}$$

To finish up, combine (27) and (29) 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. ∎

## 6. Discussion

In this work, we have investigated stepsize-based acceleration of gradient descent, with particular emphasis on achieving anytime convergence guarantees. We have designed a (predetermined) stepsize schedule that achieves a $O(1/T^{\frac{2\log_2\rho}{\log_2\rho+1}}) \approx O(T^{-1.119})$ convergence rate (with $\rho = \sqrt{2}+1$), which holds simultaneously for any stopping time $T$ without the need of knowing $T$ *a priori*. Our results have provided a positive answer to, and hence settled, the open problem proposed by Kornowski and Shamir (2024) in COLT 2024. Moving forward, it would be fundamentally important to develop lower bounds in order to assess the tightness of our convergence guarantees. It would also be of great interest to extend such anytime stepsize-based acceleration results to broader optimization algorithms like proximal gradient methods.

## Acknowledgments

Y. Chen is supported in part by the Sloan Research Fellowship, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. JDL acknowledges support of Open Philanthropy, NSF IIS-2107304, NSF CCF-2212262, ONR Young Investigator Award, NSF CAREER Award 2144994, and NSF CCF-2019844. SSD acknowledges the support of NSF IIS-2110170, NSF DMS-2134106, NSF CCF-2212261, NSF IIS-2143493, NSF CCF-2019844, NSF IIS-2229881, and the Sloan Research Fellowship.

## References

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

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

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

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Antoine Daccache, François Glineur, and Julien Hendrickx. *Performance estimation of the gradient method with fixed arbitrary step sizes*. PhD thesis, Master's thesis, Université Catholique de Louvain, 2019.

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

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

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

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

Benjamin Grimmer, Kevin Shu, and Alex L Wang. Accelerated gradient descent via long steps. *arXiv preprint arXiv:2309.09961*, 2023.

Benjamin Grimmer, Kevin Shu, and Alex L Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *arXiv preprint arXiv:2403.14045*, 2024a.

Benjamin Grimmer, Kevin Shu, and Alex L Wang. Composing optimized stepsize schedules for gradient descent. *arXiv preprint arXiv:2410.16249*, 2024b.

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

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Teodor Rotaru, François Glineur, and Panagiotis Patrinos. 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*, 2024.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

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

Zehao Zhang and Rujun Jiang. Accelerated gradient descent by concatenation of stepsize schedules. *arXiv preprint arXiv:2410.12395*, 2024.

## Appendix A. Proof of Lemma 8

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

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

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

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

$$t \leq \sum_{j=1}^{o_t-1} k_j 2^j + m 2^{o_t} \leq 4 \cdot 2^{(c+1)(o_t-1)} + m 2^{o_t};$$
$$A_{t+1}(\widehat{\boldsymbol{s}}) \geq \sum_{j=1}^{o_t-1} (\rho^j - 1) \cdot k_j + (m-1)(\rho^{o_t} - 1) \geq \frac{1}{2} \cdot 2^{(c+\log_2 \rho)(o_t-1)} + \frac{m-1}{2} \rho^{o_t},$$

where the second line invokes Lemma 7.

- If $m 2^{o_t} \leq 2^{(c+1)(o_t-1)}$, then we have $t \leq 3 \cdot 2^{(c+1)(o_t-1)}$, which means that $A_{t+1}(\widehat{\boldsymbol{s}}) \geq \frac{1}{2} \cdot 2^{(c+\log_2 \rho)(o_t-1)} \geq \frac{1}{18} t^{\frac{c+\log_2 \rho}{c+1}}$.
- If $m 2^{o_t} > 2^{(c+1)(o_t-1)}$ — i.e, $2^{o_t c} \geq m > 2^{o_t c - c - 1} \geq 1$ — then one has
$$t^{\frac{c+\log_2 \rho}{c+1}} \leq \left(4 \cdot 2^{(c+1)(o_t-1)} + m 2^{o_t}\right)^{\frac{c+\log_2 \rho}{c+1}} < 9 (m 2^{o_t})^{\frac{c+\log_2 \rho}{c+1}} \leq 9 \cdot m \rho^{o_t} \leq 36 \cdot \frac{m-1}{2} \rho^{o_t} \leq 36 A_{t+1}(\widehat{\boldsymbol{s}}).$$

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

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

$$t \geq \sum_{j=1}^{o_t-1} k_j 2^j \geq \sum_{j=1}^{o_t-1} 2^{(c+1)j} \geq 2^{(c+1)(o_t-1)}, \tag{30}$$

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

## Appendix B. Proof of Lemma 9

From Definition 3 of the primitive stepsize schedule, we obtain

$$\begin{aligned} & A_k (f_k - f^*) + C_k \|\boldsymbol{g}_k\|^2 + \frac{1}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\ \leq \quad & \frac{1}{2} \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2} \|\boldsymbol{g}_i\|^2 \right). \end{aligned} \tag{31}$$

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

$$\sum_{i=1}^{k-1} \alpha_i \left( f_i - f_0 - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_0 \rangle + \frac{1}{2} \|\boldsymbol{g}_i - \boldsymbol{g}_0\|^2 \right) \leq 0, \tag{32}$$

which further implies that

$$\begin{aligned}
&\sum_{i=1}^{k-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right) \\
&\leq \sum_{i=1}^{k-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right) - \sum_{i=1}^{k-1} \alpha_i \left( f_i - f_0 - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_0 \rangle + \frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_0\|^2 \right) \\
&= \sum_{i=1}^{k-1} \alpha_i \left( f_0 - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{1}{2}\|\boldsymbol{g}_0\|^2 \right). \qquad (33)
\end{aligned}$$

Substituting (33) into (31) and using the fact that $\sum_{i=1}^{k-1} \alpha_i \boldsymbol{g}_i = \boldsymbol{x}_1 - \boldsymbol{x}_k$, we can derive

$$\begin{aligned}
&A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \\
&\leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1} \alpha_i \left( f_0 - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{1}{2}\|\boldsymbol{g}_0\|^2 \right) \\
&= \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + A_k(f_0 - f^*) - \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2 \\
&= \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + A_k(f_0 - f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_k, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2. \qquad (34)
\end{aligned}$$

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 (34) yields

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

as claimed.

## Appendix C. Proof of Lemma 10

Because $\boldsymbol{\alpha}_{1:k-1}$ is a primitive stepsize schedule, it follows from Lemma 9 that

$$A_k(f_k - f_0) + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2 + C_k\|\boldsymbol{g}_k\|^2 \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2. \qquad (35)$$

We also make note of the following basic facts:

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha_0 \boldsymbol{g}_0; \tag{36a}$$

$$\sum_{i=1}^{k-1} \alpha_i \boldsymbol{g}_i = \boldsymbol{x}_1 - \boldsymbol{x}_k = \boldsymbol{x}_0 - \boldsymbol{x}_k - \alpha_0 \boldsymbol{g}_0; \tag{36b}$$

$$f_0 - f_k \le \langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_k \rangle - \frac{1}{2}\|\boldsymbol{g}_0 - \boldsymbol{g}_k\|^2; \tag{36c}$$

$$(A_k + 1)\langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_k \rangle \le \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2 + \frac{(A_k+1)^2}{2}\|\boldsymbol{g}_0\|^2. \tag{36d}$$

Putting the above inequalities together, we arrive at

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

(19) is proven.

To prove (20), it suffices to note from (35) that

$$\begin{aligned} A_k(f_k - f_0) + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2 &\le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \sum_{i=1}^{k-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2 \\ &= \frac{\alpha_0^2}{2}\|\boldsymbol{g}_0\|^2 + \langle \boldsymbol{x}_0 - \boldsymbol{x}_k - \alpha_0 \boldsymbol{g}_0, \boldsymbol{g}_0 \rangle - \frac{A_k}{2}\|\boldsymbol{g}_0\|^2 \\ &= \left( \frac{1}{2}\alpha_0^2 - \frac{A_k}{2} - \alpha_0 \right) \|\boldsymbol{g}_0\|^2 + \langle \boldsymbol{x}_0 - \boldsymbol{x}_k, \boldsymbol{g}_0 \rangle \\ &\le \left( \frac{1}{2}\alpha_0^2 - \frac{A_k}{2} - \alpha_0 + \frac{1}{2} \right) \|\boldsymbol{g}_0\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_0\|^2, \end{aligned} \tag{37}$$

where the second line makes use of (36a) and (36b), and the last line results from the elementary inequality $2\langle \boldsymbol{a}, \boldsymbol{b} \rangle \le \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2$. This concludes the proof.

## Appendix D. Proof of Lemma 11

Denote by $\widehat{A}_j = \rho^j - 1$ the aggregate stepsize of the $j$-th order silver stepsize schedule for $j \ge 1$ (see, e.g., Lemma 7). Consider any $\ell \in [1, k-1]$, then there exist $1 \le p \le i$ and $i > m_1 > m_2 > \dots, > m_p \ge 0$ such that

$$\ell = \sum_{j=1}^{p} 2^{m_j}.$$

Also, take

$$\tau_0 = 0 \qquad \text{and} \qquad \tau_j = \sum_{j'=1}^{j} 2^{m_j},$$

and hence $\tau_p = \ell$.

Now, consider the stepsize schedule $\boldsymbol{\alpha}_{\tau_j:\tau_{j+1}-1} = [\alpha_i]_{\tau_j \le i < \tau_{j+1}}$, whose length is $\tau_{j+1} - \tau_j = 2^{m_{j+1}}$. By construction, we know that $\boldsymbol{\alpha}_{\tau_j+1:\tau_{j+1}-1}$ corresponds to the $m_{j+1}$-th order silver stepsize schedule, and

$$\alpha_{\tau_{j+1}} = (\sqrt{2}-1)\widehat{A}_{m_{j+1}} + \sqrt{2}$$

for all $j$. Combining this with the fact $\widehat{A}_j = \rho^j - 1$ and the assumption that $\alpha \geq (\sqrt{2}-1)A_k + \sqrt{2}$ (recall that $\alpha = \alpha_0$) yields

$$\alpha_{\tau_{j+1}} \leq \alpha_{\tau_j}/2, \tag{38}$$

provided that $j \geq 0$ and $m_{j+1} \geq 2$.

Invoking Lemma 10 with this stepsize schedule, we can demonstrate that

$$\widehat{A}_{m_{j+1}}(\widehat{A}_{m_{j+1}}+1)\|\boldsymbol{g}_{\tau_{j+1}}\|^2 \leq \left(\alpha_{\tau_j}^2 - 2\alpha_{\tau_j} + \widehat{A}_{m_{j+1}}^2 + \widehat{A}_{m_{j+1}} + 1\right)\|\boldsymbol{g}_{\tau_j}\|^2; \tag{39}$$

$$f_{\tau_{j+1}} - f_{\tau_j} \leq \frac{1}{\widehat{A}_{m_{j+1}}}\left(\frac{1}{2}\alpha_{\tau_j}^2 - \frac{\widehat{A}_{m_{j+1}}}{2} - \alpha_{\tau_j} + 1\right)\|\boldsymbol{g}_{\tau_j}\|^2 \leq \frac{1}{2}\alpha_{\tau_j}^2\|\boldsymbol{g}_{\tau_j}\|^2. \tag{40}$$

It then follows from (39) that

$$\begin{aligned}\|\boldsymbol{g}_{\tau_{j+1}}\|^2 &\leq \frac{\leq \left(\alpha_{\tau_j}^2 + \widehat{A}_{m_{j+1}}^2 + \widehat{A}_{m_{j+1}} + 1\right)\|\boldsymbol{g}_{\tau_j}\|^2}{\widehat{A}_{m_{j+1}}(\widehat{A}_{m_{j+1}}+1)} \\ &\leq \left(\frac{\alpha_{\tau_j}^2}{\widehat{A}_{m_{j+1}}(\widehat{A}_{m_{j+1}}+1)} + 1\right)\|\boldsymbol{g}_{\tau_j}\|^2, \end{aligned} \tag{41}$$

which we would like further control by dividing into two cases.

- *Case 1:* $m_{j+1} \geq 2$. In this case, we have $\widehat{A}_{m_{j+1}} \geq \rho^2 - 1 = 2 + 2\sqrt{2}$. Observing that $\alpha_{\tau_{j+1}} = (\sqrt{2}-1)\widehat{A}_{m_{j+1}} + \sqrt{2}$ by construction, one can easily verify that

  $$\widehat{A}_{m_{j+1}}(\widehat{A}_{m_{j+1}}+1) \geq (\sqrt{2}+1)\alpha_{\tau_{j+1}}^2,$$

  which combined with (41) implies that

  $$\|\boldsymbol{g}_{\tau_{j+1}}\|^2 \leq \left(\frac{\alpha_{\tau_j}^2}{\alpha_{\tau_{j+1}}^2}\cdot(\sqrt{2}-1) + 1\right)\|\boldsymbol{g}_{\tau_j}\|^2. \tag{42}$$

  This taken together with the property $\alpha_{\tau_{j+1}} \leq \alpha_{\tau_j}/2$ (cf. (38)) leads to

  $$\alpha_{\tau_{j+1}}^2\|\boldsymbol{g}_{\tau_{j+1}}\|^2 \leq \left(\sqrt{2} - \frac{3}{4}\right)\alpha_{\tau_j}^2\|\boldsymbol{g}_{\tau_j}\|^2. \tag{43}$$

- *Case 2:* $m_{j+1} < 2$. In this case, it is readily seen from (41) that

  $$\|\boldsymbol{g}_{\tau_{j+1}}\|^2 \leq \left(\frac{\alpha_{\tau_j}^2}{\widehat{A}_{m_{j+1}}(\widehat{A}_{m_{j+1}}+1)} + 1\right)\|\boldsymbol{g}_{\tau_j}\|^2 \leq \alpha_{\tau_j}^2\|\boldsymbol{g}_{\tau_j}\|^2.$$

  Moreover, we make the observation that

  $$\alpha_{\tau_{j+1}} = (\sqrt{2}-1)\widehat{A}_{m_{j+1}} + \sqrt{2} \leq (\sqrt{2}-1)(\rho-1) + \sqrt{2} = 2,$$

  which allows us to reach

  $$\alpha_{\tau_{j+1}}^2\|\boldsymbol{g}_{\tau_{j+1}}\|^2 \leq 12\alpha_{\tau_j}^2\|\boldsymbol{g}_{\tau_j}\|^2. \tag{44}$$

Putting (43) and (44) together, we can conclude that for any $j \geq 1$,

$$\alpha_{\tau_j}^2 \|\boldsymbol{g}_{\tau_j}\|^2 \leq 432 \left(\sqrt{2} - \frac{3}{4}\right)^j \alpha_{\tau_0}^2 \|\boldsymbol{g}_{\tau_0}\|^2 = 432 \left(\sqrt{2} - \frac{3}{4}\right)^j \alpha^2 \|\boldsymbol{g}_0\|^2. \tag{45}$$

This taken collectively with (40) gives

$$\begin{aligned} f_\ell - f_0 &= \sum_{j=0}^{p-1} (f_{\tau_{j+1}} - f_{\tau_j}) \leq \frac{1}{2} \sum_{j=0}^{p-1} \alpha_{\tau_j}^2 \|\boldsymbol{g}_{\tau_j}\|^2 \\ &\leq \left(\frac{1}{2} + 216 \sum_{j\geq 1} \left(\sqrt{2} - \frac{3}{4}\right)^j\right) \alpha^2 \|\boldsymbol{g}_0\|^2 \leq 432\alpha^2 \|\boldsymbol{g}_0\|^2 \end{aligned} \tag{46}$$

as claimed.

## Appendix E. Proof of preliminary facts from Zhang and Jiang (2024)

### E.1. Proof of Lemma 4

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 4. We shall provide a proof in Appendix E.2.

**Lemma 13** *(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*

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

Next, in view of the definition of the primitive stepsize schedule (cf. Definition 3), we can easily see that

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

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

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

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

Lemma 13 tells us that

$$(x + \alpha)(f_k - f_\ell) \leq -\frac{(x + \alpha)(y + 3\alpha - 2\alpha^2)}{2(y + 2 - \alpha)} \|\boldsymbol{g}_\ell\|^2 - \frac{(x + \alpha)(2y^2 + 3y + \alpha)}{2(y + 2 - \alpha)} \|\boldsymbol{g}_k\|^2. \tag{47}$$

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

$$L_1 + L_2 + L_3 + L_4 \le R_1 + R_2 + R_3 + R_4, \tag{48}$$

where

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

We now proceed to simplify (48). 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 (48) then results in $L_1 \le R_1$, namely,

$$\begin{aligned} A_k(f_k - f^*) + C_k\|\boldsymbol{g}_k\|^2 + \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 &\le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{k-1}\alpha_i\left(f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^*\rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2\right), \\ &\le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2, \end{aligned} \tag{49}$$

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

### E.2. Proof of Lemma 13

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

$$A_\ell(f_\ell - f^*) + C_\ell \|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\|^2 \le \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \sum_{i=1}^{\ell-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle) + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right). \tag{50}$$

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

$$\begin{aligned} & \sum_{i=1}^{\ell-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right) \\ &\quad \le A_\ell(f_\ell - f^*) - \sum_{i=1}^{\ell} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{x}_\ell - \boldsymbol{x}^* \rangle + \sum_{i=1}^{\ell-1} \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{g}_\ell \rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_\ell\|^2 \\ &\quad = A_\ell(f_\ell - f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_\ell - \boldsymbol{x}^* \rangle + \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{g}_\ell \rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_\ell\|^2, \end{aligned} \tag{51a}$$

and similarly,

$$\begin{aligned} & \sum_{i=1}^{\ell-1} \alpha_i \left( f_i - f^* - \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}^* \rangle + \frac{1}{2}\|\boldsymbol{g}_i\|^2 \right) \\ &\quad \le A_\ell(f_0 - f^*) - \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \langle \boldsymbol{x}_1 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 \rangle - \frac{A_\ell}{2}\|\boldsymbol{g}_0\|^2. \end{aligned} \tag{51b}$$

Combine (51a) and (51b) 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 (50), 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

$$
\begin{aligned}
&A_\ell(f_0 - f_\ell)\\
&\ge 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_1 - \boldsymbol{x}_\ell\rangle - \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2\\
&= 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \alpha\boldsymbol{g}_0 - \boldsymbol{x}_\ell\rangle - \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2\\
&= 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \alpha\langle \boldsymbol{g}_0, \boldsymbol{x}_0 - \boldsymbol{x}_\ell\rangle - \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, \boldsymbol{g}_0 + \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2 - \alpha^2\|\boldsymbol{g}_0\|^2\\
&= 2C_\ell\|\boldsymbol{g}_\ell\|^2 + \frac{1}{2}A_\ell(\|\boldsymbol{g}_\ell\|^2 + \|\boldsymbol{g}_0\|^2) + \langle \boldsymbol{x}_0 - \boldsymbol{x}_\ell, (\alpha - 1)\boldsymbol{g}_0 - \boldsymbol{g}_\ell\rangle + \alpha\langle \boldsymbol{g}_0, \boldsymbol{g}_\ell\rangle + \alpha\|\boldsymbol{g}_0\|^2 - \alpha^2\|\boldsymbol{g}_0\|^2.
\end{aligned} \tag{52}
$$

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

$$\langle \boldsymbol{x}_0-\boldsymbol{x}_\ell, (\alpha-1)\boldsymbol{g}_0-\boldsymbol{g}_\ell\rangle+\alpha\langle \boldsymbol{g}_0,\boldsymbol{g}_\ell\rangle \ge (\alpha-2)(f_0-f_\ell)+\frac{\alpha}{2}(\|\boldsymbol{g}_0\|^2+\|\boldsymbol{g}_\ell\|^2). \tag{53}$$

Substitution into (52) then leads to

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

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

$$(A_\ell+2-\alpha)(f_0-f_\ell) \ge \left(A_\ell^2+\frac{3A_\ell}{2}+\frac{\alpha}{2}\right)\|\boldsymbol{g}_\ell\|^2+\left(\frac{A_\ell}{2}+\frac{\alpha}{2}+\alpha-\alpha^2\right)\|\boldsymbol{g}_0\|^2. \tag{54}$$

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