Gradient Descent's Last Iterate is Often (slightly) Suboptimal

Guy Kornowski
Weizmann Institute, Institute for Advanced Study
gkornowski@ias.edu

Ohad Shamir
Weizmann Institute, University of Toronto, Vector Institute
ohad.shamir@weizmann.ac.il

Abstract

We consider the well-studied setting of minimizing a convex Lipschitz function using either gradient descent (GD) or its stochastic variant (SGD), and examine the last iterate convergence. By now, it is known that standard stepsize choices lead to a last iterate convergence rate of $\log T/\sqrt{T}$ after T steps. A breakthrough result of Jain et al. [2019] recovered the optimal $1/\sqrt{T}$ rate by constructing a non-standard stepsize sequence. However, this sequence requires choosing T in advance, as opposed to common stepsize schedules which apply for any time horizon. Moreover, Jain et al. conjectured that without prior knowledge of T, no stepsize sequence can ensure the optimal error for SGD's last iterate, a claim which so far remained unproven. We prove this conjecture, and in fact show that even in the noiseless case of GD, it is impossible to avoid an excess poly-log factor in T when considering an anytime last iterate guarantee.

1 Introduction

Let $f : \mathcal{X} \to \mathbb{R}$ be a convex Lipschitz function over a convex domain $\mathcal{X} \subset \mathbb{R}^d$, and consider the stochastic gradient descent (SGD) algorithm[1] starting from $\mathbf{x}_0 \in \mathcal{X}$:

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{X}}\left(\mathbf{x}_t - \eta_t \mathbf{g}_t\right) \quad \text{for all} \quad t \in \mathbb{N}\,, \tag{1}$$

where $\mathbb{E}[\mathbf{g}_t \mid \mathbf{x}_t] \in \partial f(\mathbf{x}_t)$, and $(\eta_0, \eta_1, \dots)$ is the stepsize sequence (also referred to as learning rate).

Although SGD has been studied extensively for decades [Robbins and Monro, 1951, Nemirovski and Yudin, 1983], and is commonly considered the main workhorse of machine learning [Zhang, 2004, Bottou, 2010, Goodfellow et al., 2016], there still remain some surprisingly fundamental gaps in its theoretical analysis.

The textbook analysis of SGD implies that for stepsizes $\eta_t = \Theta(1/\sqrt{t})$, the error of the *average* iterate ($\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$ after T steps) converges at a $1/\sqrt{T}$ rate (cf. Hazan, 2016, Theorem 3.4). This well-known result has the clear drawback that in practice, it is much more common to return the *last* iterate obtained, namely $\mathbf{x}_T$, rather than the average iterate. However, the last iterate of SGD with the aforementioned stepsizes only converges at a suboptimal $\log(T)/\sqrt{T}$ rate [Shamir and Zhang, 2013, Harvey et al., 2019], differing from the information-theoretically optimal $1/\sqrt{T}$ rate by an excess log factor.

[1] Strictly speaking, in this work we do not assume differentiability of f, and consider *sub*-gradient methods.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This unsatisfying aspect led to substantial research efforts analyzing the last iterate convergence of SGD under various settings, as well as motivated the design of modified output rules that recover the optimal rate. We further discuss prior works along these lines in Section 2.

Ultimately, Jain et al. [2019] proved that the last iterate of SGD, with suitable stepsizes, can indeed converge at an optimal $1/\sqrt{T}$ rate after T steps. At the heart of this remarkable result is the design of a non-trivial stepsize sequence $(\eta_t)_{t=0}^{T-1}$, which unfortunately still suffers from a shortcoming: Its values depend on T, and therefore the number of steps T must be chosen in advance. In other words, after completing a predetermined budget of T steps, if additional steps are required or desired, it is unclear how to continue minimizing the error at an optimal rate. However, in many applications, it is desirable to have algorithms with anytime guarantees, and the anytime nature of SGD with standard stepsizes is arguably one of its appeals in the first place. This leads to an undesirable mismatch between theory and practice, which affects the practical design of learning rate schedules without prior knowledge of the time horizon (see for example Defazio et al. [2024]).

The described issue was raised already by Jain et al. [2019], where the authors conjectured that

> *"in absence of a priori information about T, no step-size sequence can ensure the information theoretically optimal error rates for the final iterate of SGD."*

In other words, the conjecture states that in order to converge at an optimal rate, SGD requires either averaging, or knowing T in advance. Jain et al. provided some evidence for the validity of their conjecture by considering noisy gradient oracles and their probabilistic consequences, however, the conjecture has so far remained unproven. This leaves a natural open problem in the theoretical analysis of SGD, as to whether the last iterate can achieve optimal convergence in an anytime fashion.

In this work, we resolve this question by answering it negatively. In fact, we show that even in the *noiseless* case, the last iterate of gradient descent (GD, i.e., $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ deterministically) cannot possibly avoid an excess poly-log factor compared to the optimal rate, whenever the stopping time is not carefully chosen in advance. This stronger lower bound for anytime GD proves the conjecture of Jain et al. [2019] for SGD as a special case.

The paper is structured as follows. After discussing related work in Section 2, we present in Section 3 the formal setting that we consider. In Section 4 we state our main result, and discuss some notable consequences. In Section 5 we present the proof, besides some technical details which are deferred to the appendix. We conclude in Section 6.

2 Related Work

Last iterate of SGD. As previously discussed, classical analyses of SGD typically deal with the convergence of the average iterate [Polyak and Juditsky, 1992, Nemirovski et al., 2009]. Zhang [2004] analyzed the convergence rate of the last iterate of SGD with constant stepsize for learning linear predictors. This was extended to general convex objectives and decaying stepsizes in Shamir and Zhang [2013]. Harvey et al. [2019] further established tight high probability bounds. Tight constants were derived for GD in the deterministic setting by Zamani and Glineur [2023]. The general smooth setting was studied by Moulines and Bach [2011], with results later improved by Taylor and Bach [2019], Liu and Zhou [2024b]. Several recent works studied last iterate convergence in the so called (near-)interpolation or low-noise regime, first for least-squares [Varre et al., 2021] and subsequently for general smooth losses [Attia et al., 2025, Garrigos et al., 2025]. Several works considered objectives with an empirical risk minimization structure, and established last iterate convergence for SGD with respect to different sampling schemes [Gower et al., 2019, Liu and Zhou, 2024a], which was further studied through connections to continual learning [Evron et al., 2025, Cai and Diakonikolas, 2025].

Optimal convergences rates. Several works considered modifications of SGD or of its output rule in order to remove excess log factors and recover optimal rates for strongly-convex objectives [Hazan and Kale, 2014, Rakhlin et al., 2012, Lacoste-Julien et al., 2012]. The question of whether some form of averaging is needed for SGD to recover optimal rates in this context was raised by Shamir [2012], and answered affirmatively by Harvey et al. [2019] both for strongly-convex and convex objectives. However, their results applied only for the theoretically standard choice of step sizes ($\Theta(1/\sqrt{t})$ in the

convex case and $\Theta(1/t)$ in the strongly-convex case). As previously discussed, Jain et al. [2019] then showed that it is possible to recover optimal rates with SGD's last iterate by considering non-standard stepsizes that depend on the stopping time T.

Anytime smooth GD. In the related setting of smooth convex deterministic optimization, recent works designed non-standard stepsizes for GD that accelerate the well-established convergence rates of constant stepsizes [Altschuler and Parrilo, 2024, Grimmer et al., 2025]. It was noted by Kornowski and Shamir [2024] that these results do not yield anytime improvements, and it was asked whether such improvements are possible, which was then answered positively by Zhang et al. [2025]. It is interesting to note that in the smooth deterministic setting, the best known anytime bounds differ from non-anytime bounds by polynomial factors, and it is an open problem as to whether this is necessary. In contrast, in this work we prove that poly-logarithmic gaps are necessary in the non-smooth setting.

3 Preliminaries

Notation. We denote $\mathbb{N} := \{1, 2, \dots\}$, $\mathbb{N}_0 := \{0, 1, 2, \dots\}$, $\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} : x \geq 0\}$ and $[n] := \{1, \dots, n\}$. We use bold-faced font to denote vectors, e.g., $\mathbf{x} \in \mathbb{R}^d$, and denote by $\|\cdot\|$ the Euclidean norm. We denote $\mathrm{dist}(\mathbf{x}, A) := \inf_{\mathbf{a} \in A} \|\mathbf{x} - \mathbf{a}\|$ for $A \subset \mathbb{R}^d$, $\mathbf{x} \in \mathbb{R}^d$, and by Proj_A the projection operator onto A. $\partial f(\cdot)$ denotes the sub-gradient set of a convex function f. A function $f : \mathcal{X} \to \mathbb{R}$ is called G-Lipschitz if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X} : |f(\mathbf{x}) - f(\mathbf{y})| \leq G \|\mathbf{x} - \mathbf{y}\|$, or equivalently for a convex function, for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{g} \in \partial f(\mathbf{x}) : \|\mathbf{g}\| \leq G$. We use the standard big-O and little-o notation, with $O(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ hiding absolute constants that do not depend on problem parameters, and $a_t = o(b_t)$ meaning $\lim_{t \to \infty} a_t / b_t = 0$.

Setting. Throughout the paper we impose the standard assumptions that $f : \mathcal{X} \to \mathbb{R}$ is convex and G-Lipschitz where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex domain. Given an initial point $\mathbf{x}_0 \in \mathcal{X}$, we consider the sub-gradient method (as in Eq. (1)) so that $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$, with stepsize sequence $\boldsymbol{\eta} = (\eta_t)_{t=0}^{\infty} \in \mathbb{R}_{\geq 0}^{\mathbb{N}_0}$.

Given a stepsize sequence, we are interested in its worst-case anytime last iterate guarantee:

Definition 1. *For stepsize sequence $\boldsymbol{\eta} = (\eta_t)_{t=0}^{\infty}$, we say that $\mathcal{E}_{\boldsymbol{\eta}}^{G,D} : \mathbb{N} \to \mathbb{R}_{\geq 0}$ is an* anytime convergence rate guarantee *for $\boldsymbol{\eta}$, if for all G-Lipschitz and convex functions f over a domain $\mathcal{X}$ of diameter at most D, and initialization $\mathbf{x}_0 \in \mathcal{X}$, it holds that*

$$f(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \leq \mathcal{E}_{\boldsymbol{\eta}}^{G,D}(t) \quad \textit{for all} \quad t \in \mathbb{N} \,.$$

For example, we recall the following known anytime last iterate convergence guarantee:

Example 2 ([Shamir and Zhang, 2013])**.** *The stepsize sequence $\boldsymbol{\eta} = (D/G\sqrt{t+1})_{t=0}^{\infty}$ has the anytime convergence guarantee*

$$\mathcal{E}_{\boldsymbol{\eta}}^{G,D}(t) = \frac{DG(4 + 2\log t)}{\sqrt{t}} \,.$$

4 Main Result

We are now ready to present our main result:

Theorem 3. *No stepsize sequence $\boldsymbol{\eta} \in \mathbb{R}_{\geq 0}^{\mathbb{N}_0}$ has an anytime convergence guarantee satisfying*

$$\mathcal{E}_{\boldsymbol{\eta}}^{G,D}(t) = o\left(\frac{DG \log^{1/12}(t)}{\sqrt{t}}\right) \quad \textit{as } t \to \infty \,.$$

In particular, neither SGD nor GD's last iterate can yield the optimal $1/\sqrt{t}$ rate in an anytime fashion.

We remark that the constant $1/12$ that appears in the logarithm's exponent was not carefully optimized.

Remark 4 (Absence of noise). *As previously mentioned, Theorem 3 confirms the conjecture of Jain et al. [2019], and is in fact stronger as it applies even in the noiseless case of GD. It is interesting to note that an intuition conveyed by Jain et al. [2019] is that noise in the sub-gradients can lead SGD to be "bad in expectation" (precisely defined therein) even in one dimension. Our proof is based on a different perspective, where instead of noise, the key factors driving the lower bound constructions are high-dimensionality, and the stepsize sequence not "knowing" in advance when the algorithm should stop. We discuss this in more detail in Section 5.*

Before delving into the analysis in the next section, we present here the high-level idea behind the proof of Theorem 3. As mentioned in the introduction, Harvey et al. [2019] showed that any particular stepsize schedule of the form $\eta_t = c/\sqrt{t}$ (with $c > 0$ a fixed absolute constant) leads to an extra log factor in the last iterate error. We therefore aim at showing that in a sense, any successful stepsize sequence cannot be too far from this form, and thus leads to a similar (albeit slightly looser) bound. Specifically, in Lemmas 5 and 6 we establish that stepsizes that lead to $\tilde{O}(1/\sqrt{t})$ convergence, up to at most a log factor, must satisfy on one hand that $\eta_t = \tilde{O}(1/\sqrt{t})$ at every step, but also that $\sum_{j=0}^{t} \eta_j = \tilde{\Omega}(\sqrt{t})$ for every cumulative sum. Thus, by examining sufficiently long time intervals, we can deduce that any stepsize sequence $\boldsymbol{\eta}$ that leads to $\tilde{O}(1/\sqrt{t})$ last iterate error, cannot be "too far" from $\eta_t \approx 1/\sqrt{t}$ up to log factors at "most" time indices. Then, in Lemma 8 we provide a robust extension of the lower bound construction used in the analysis of Harvey et al. [2019], which instead of assuming that $\eta_t = c/\sqrt{t}$ exactly, depends directly on the values of $\boldsymbol{\eta}$. By combining these pieces, further algebraic manipulations result in our lower bound.

5 Proof of Theorem 3

In this section, we provide the core arguments that are used to prove Theorem 3. Additional technical proofs of lemmas are deferred to the appendix.

First, by a rescaling argument, which replaces any G-Lipschitz function f over any domain $\mathcal{X}$ of diameter D by $\tilde{f}(\mathbf{x}) = \frac{2}{GD} f(\frac{D}{2}\mathbf{x})$ (which is 1-Lipschitz over the domain $\frac{2}{D}\mathcal{X}$ of diameter 2), we see that it suffices to prove Theorem 3 for $G = 1$, $D = 2$. We abbreviate notation accordingly by denoting $\mathcal{E}_{\boldsymbol{\eta}} = \mathcal{E}^{1,2}_{\boldsymbol{\eta}}$.

Let $\boldsymbol{\eta} = (\eta_t)_{t=0}^{\infty}$ be some stepsize sequence, and suppose it has a corresponding anytime guarantee $\mathcal{E}_{\boldsymbol{\eta}}$ satisfying

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \leq \frac{\phi(t)}{\sqrt{t}} \ , \quad \text{where } \ \phi : \mathbb{N} \to [1, \infty) \text{ is non-decreasing.}$$

For instance, ϕ might be of the form $\phi(t) = c_1 \log^{c_2}(t) + c_3$ for some constants $c_1, c_2, c_3 > 0$. Our goal is to show that it *cannot* be that $\phi(t) = o(\log^{1/12}(t))$.

We start by establishing basic lower bounds on $\mathcal{E}_{\boldsymbol{\eta}}(t)$, and therefore on $\phi(t)$ (which is at least $\mathcal{E}_{\boldsymbol{\eta}}(t)\sqrt{t}$), in terms of the stepsizes $(\eta_0, \ldots, \eta_{t-1})$. The first such bound shows that large steps can result in large last iterate error (in the worst-case), in case the algorithm happens to stop right after a large step. Formally, we show the following:

Lemma 5. *For any stepsize sequence $\boldsymbol{\eta}$ and $t \in \mathbb{N}$, it holds that*

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \geq \min\{\eta_{t-1}, 1\} \ .$$

Proof of Lemma 5. Given $\boldsymbol{\eta}$ and t, let $\epsilon > 0$ be some arbitrarily small number so that $\epsilon < \min\{1, \eta_{t-1}, \sum_{j=0}^{t-2} \eta_j\}$, and denote $c_\epsilon := \frac{\epsilon}{\sum_{j=0}^{t-2} \eta_j}$. Consider the univariate function over $[-1, 1]$:

$$f(x) = \begin{cases} -x, & \text{if } x < 0 \\ c_\epsilon x, & \text{if } 0 \leq x \leq \epsilon \\ x - \epsilon + c_\epsilon \epsilon, & \text{if } x > \epsilon \end{cases} .$$

Note that f is 1-Lipschitz, convex since $c_\epsilon < 1$, and minimized at $f(0) = 0$. Consider the iterates of GD initialized at $x_0 = \epsilon$. It is easy to see that for $i < t - 1$ it holds that $x_{i+1} = x_i - \eta_i c_\epsilon$, and

therefore

$$x_{t-1} = \epsilon - c_\epsilon \sum_{j=0}^{t-2} \eta_j = 0 \,.$$

Suppose that at time $t-1$ the sub-gradient is given by $-1 \in \partial f(x_{t-1})$, thus resulting in $x_t = \mathrm{Proj}_{[-1,1]}(\eta_{t-1}) = \min\{\eta_{t-1}, 1\}$. Noting that $\eta_{t-1} > \epsilon$, we get that $\mathcal{E}_{\boldsymbol{\eta}}(t) \geq f(x_t) = \min\{\eta_{t-1}, 1\} - \epsilon + c_\epsilon \epsilon$. Since this inequality holds for arbitrarily small ϵ, it must hold that

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \geq \lim_{\epsilon \to 0^+} (\min\{\eta_{t-1}, 1\} - \epsilon + c_\epsilon \epsilon) = \min\{\eta_{t-1}, 1\} \,.$$

□

Second, we prove a lower bound inversely proportional to the accumulated stepsize sum:

Lemma 6. *For any stepsize sequence $\boldsymbol{\eta}$ and $t \in \mathbb{N}$ such that $\sum_{j=0}^{t-1} \eta_j \geq \frac{1}{2}$, it holds that*

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \geq \frac{1}{4e^2 \sum_{j=0}^{t-1} \eta_j} \,.$$

Proof of Lemma 6. The proof closely follows the proof of Kornowski and Shamir [2024, Theorem 1]. Given $\boldsymbol{\eta}$ and t, consider the convex quadratic

$$f(x) = \frac{x^2}{4 \sum_{j=0}^{t-1} \eta_j} \,.$$

Note that f is 1-Lipschitz over $[-1, 1]$ as long as $\sum_{j=0}^{t-1} \eta_j \geq \frac{1}{2}$, and is minimized at $f(0) = 0$. Examining the iterates of GD initialized at $x_0 = 1$, we get that

$$x_{i+1} = x_i - \eta_i \cdot \nabla f(x_i) = x_i - \eta_i \cdot \frac{x_i}{2 \sum_{j=0}^{t-1} \eta_j} = \left(1 - \frac{\eta_i}{2 \sum_{j=0}^{t-1} \eta_j}\right) x_i \,,$$

and thus a simple induction reveals that

$$x_t = \prod_{j=0}^{t-1} \left(1 - \frac{\eta_j}{2 \sum_{j=0}^{t-1} \eta_j}\right) \,.$$

Therefore, using the simple inequality $\log(1-z) \geq -2z$ for any $z \in [0, \frac{1}{2}]$, we get

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\eta}}(t) \geq f(x_t) &= \frac{1}{4 \sum_{j=0}^{t-1} \eta_j} \cdot \prod_{j=0}^{t-1} \left(1 - \frac{\eta_j}{2 \sum_{j=0}^{t-1} \eta_j}\right)^2 \\
&= \frac{1}{4 \sum_{j=0}^{t-1} \eta_j} \cdot \exp\left[2 \cdot \sum_{j=0}^{t-1} \log\left(1 - \frac{\eta_j}{2 \sum_{j=0}^{t-1} \eta_j}\right)\right] \\
&\geq \frac{1}{4 \sum_{j=0}^{t-1} \eta_j} \cdot \exp\left[-4 \cdot \sum_{j=0}^{t-1} \frac{\eta_j}{2 \sum_{j=0}^{t-1} \eta_j}\right] \\
&= \frac{e^{-2}}{4 \sum_{j=0}^{t-1} \eta_j} \,.
\end{aligned}$$

□

By considering both Lemma 5 and 6, we see that in order to achieve low anytime last iterate error, each individual step cannot be too large, yet the step sum has to grow at roughly a $\sqrt{t}$ rate up to possible log factors. In particular, combining these lemmas yields useful bounds on sums over intervals:

Lemma 7. *For any $t_1 < t_2 \in \mathbb{N}$ such that $\mathcal{E}(t) < 1$ on $t \in [t_1, t_2]$, it holds that*

$$\sum_{j=t_1}^{t_2} \eta_j \leq 2\phi(t_2+1)(\sqrt{t_2+1} - \sqrt{t_1}) \,.$$

If t_2 is such that $\sum_{j=0}^{t_2} \eta_j \geq \frac{1}{2}$, it further holds that

$$\sum_{j=t_1}^{t_2} \eta_j \geq \frac{\sqrt{t_2+1}}{4e^2\phi(t_2+1)} - 2\phi(t_1)\sqrt{t_1} \,.$$

Proof of Lemma 7. For the upper bound, by Lemma 5 and the monotonicity of ϕ we get

$$\begin{aligned}
\sum_{j=t_1}^{t_2} \eta_j &\leq \sum_{j=t_1}^{t_2} \mathcal{E}_{\boldsymbol{\eta}}(j+1) \\
&\leq \sum_{j=t_1+1}^{t_2+1} \frac{\phi(j)}{\sqrt{j}} \\
&\leq \phi(t_2+1) \int_{t_1}^{t_2+1} \frac{1}{\sqrt{x}} dx \\
&= 2\phi(t_2+1)(\sqrt{t_2+1} - \sqrt{t_1}) \,.
\end{aligned}$$

For the lower bound, by Lemmas 5 and 6, and the monotonicity of ϕ, it holds that

$$\begin{aligned}
\sum_{k=t_1}^{t_2} \eta_k &= \sum_{k=0}^{t_2} \eta_k - \sum_{k=0}^{t_1-1} \eta_k \\
&\geq \frac{1}{4e^2\mathcal{E}_{\boldsymbol{\eta}}(t_2+1)} - \sum_{k=0}^{t_1-1} \mathcal{E}_{\boldsymbol{\eta}}(k+1) \\
&\geq \frac{\sqrt{t_2+1}}{4e^2\phi(t_2+1)} - \phi(t_1) \sum_{k=1}^{t_1} \frac{1}{\sqrt{k}} \\
&\geq \frac{\sqrt{t_2+1}}{4e^2\phi(t_2+1)} - 2\phi(t_1)\sqrt{t_1} \,.
\end{aligned}$$

□

Next, we establish a lower bound on $\mathcal{E}_{\boldsymbol{\eta}}(t)$ directly in terms of the function ϕ and the stepsizes, which will later use in order to later derive a lower bound on ϕ.

Lemma 8. *For any stepsize sequence $\boldsymbol{\eta}$, any $t \in \mathbb{N}$ and ϕ as in Eq. (2), it holds that*

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \geq \frac{1}{64\phi(t+1)\sqrt{t+1}} \sum_{j=0}^{t-1} \frac{\min\{1, \eta_j\sqrt{t+1}\}^2}{(t+1-j)} \,. \tag{2}$$

The proof of Lemma 8 is substantially more involved than the previously discussed bounds. It is high-dimensional in nature, and is based on modifying a technique of Harvey et al. [2019] which constructs a lower bound for max-of-linear functions in the style of Nemirovski and Yudin [1983]. To do so, we carefully construct a sequence of linear functions $\mathbf{x} \mapsto \mathbf{v}_i^\top \mathbf{x}$ that depends both on the stepsize sequence $\boldsymbol{\eta}$ and on ϕ, such that applying GD to $\mathbf{x} \mapsto \max_i\{\mathbf{v}_i^\top \mathbf{x}\}$ results in an analytically solvable trajectory (Eq. (6) in the appendix), with a last iterate error that we can lower bound as in (2). The full details are appear in the appendix.

Following Lemma 8, we further note that $\min\{1, \eta_j\sqrt{j+1}\} \geq \eta_j\sqrt{j+1}/\phi(j+1)$, since by Lemma 5 and Eq. (2) it holds that $\phi(j+1) \geq \eta_j\sqrt{j+1}$ and $\phi(j+1) \geq 1$. Hence, Eq. (2) implies

that for any $t \in \mathbb{N}$:

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \geq \frac{1}{64\phi(t+1)\sqrt{t+1}} \sum_{j=0}^{t-1} \frac{(\eta_j \sqrt{j+1}/\phi(j+1))^2}{(t+1-j)}$$
$$\geq \frac{1}{64\phi(t+1)^3\sqrt{t+1}} \sum_{j=0}^{t-1} \frac{\eta_j^2(j+1)}{(t+1-j)} \,.$$

Therefore, since $\mathcal{E}_{\boldsymbol{\eta}}(t) \leq \frac{\phi(t)}{\sqrt{t}} \leq \frac{\phi(t+1)}{\sqrt{t}}$, we can rearrange and get

$$\phi(t+1)^4 \geq \frac{1}{128} \sum_{j=0}^{t-1} \frac{\eta_j^2(j+1)}{(t+1-j)} \geq \frac{1}{128} \sum_{j=0}^{t-1} \frac{\eta_j^2 j}{(t+1-j)} \,. \tag{3}$$

With a lower bound on ϕ in hand, it remains to lower bound the sum above. As the bound above holds for all t, we let $T \in \mathbb{N}$ be some arbitrarily large time scale which we can average over, and get that

$$\begin{aligned}
128\phi(T+1)^4 &\geq \frac{1}{T} \sum_{t=1}^{T} 128\phi(t+1)^4 \\
&\geq \frac{1}{T} \sum_{t=1}^{T} \sum_{j=0}^{t-1} \frac{j\eta_j^2}{(t+1-j)} \\
&= \frac{1}{T} \sum_{k=1}^{T-1} k\eta_k^2 \sum_{t=1}^{T} \sum_{j=0}^{t-1} \frac{\mathbb{1}\{j=k\}}{(t+1-j)} \\
&= \frac{1}{T} \sum_{k=1}^{T-1} k\eta_k^2 \sum_{t=k}^{T} \frac{1}{(t+1-k)} \\
&= \frac{1}{T} \sum_{k=1}^{T-1} k\eta_k^2 (H_{T+1-k} - 1) \\
&\geq \frac{H_{T/2} - 1}{T} \sum_{k=1}^{T/2} k\eta_k^2 \,,
\end{aligned} \tag{4}$$

where we denoted the harmonic sum $H_n := \sum_{i=1}^{n} \frac{1}{i}$.

We thus turn to lower bound the sum $\sum_{k=1}^{T/2} k\eta_k^2$. For $t_1 < T/2 - 1$ soon to be chosen it holds that

$$\begin{aligned}
\sum_{k=1}^{T/2} k\eta_k^2 &\geq \sum_{k=t_1}^{T/2} k\eta_k^2 \\
&\geq t_1 \sum_{k=t_1}^{T/2} \eta_k^2 \\
&\overset{(1)}{\geq} \frac{t_1}{T/2 - t_1 + 1} \left(\sum_{k=t_1}^{T/2} \eta_k \right)^2 \\
&\overset{(2)}{\geq} \frac{t_1}{T/2 - t_1 + 1} \left(\frac{\sqrt{T/2+1}}{4e^2\phi(T/2+1)} - 2\phi(t_1)\sqrt{t_1} \right)^2 \,,
\end{aligned} \tag{5}$$

where (1) follows from a standard ℓ_1/ℓ_2 inequality (for every vector $\mathbf{x} \in \mathbb{R}^n$: $\|\mathbf{x}\|_2^2 \geq \|\mathbf{x}\|_1^2/n$), and (2) is due to Lemma 7.[2]

[2]Note that to apply Lemma 7 we can assume that T is sufficiently large so that $\sum_{j=0}^{T} \eta_j > \frac{1}{2}$, which can be chosen as long as $\sum_{j=0}^{\infty} \eta_j > \frac{1}{2}$. If $\sum_{j=0}^{\infty} \eta_j \leq \frac{1}{2}$ on the other hand, Theorem 3 is trivial, since GD does not even converge for $f(x) = x,\ \mathcal{X} = [-1, 1],\ x_0 = 1$, let alone at the prescribed rate.

By setting

$$t_1 := \left\lfloor \frac{T/2+1}{2^8 e^4 \phi(T/2+1)^4} \right\rfloor - 1$$

it holds that

$$\frac{\sqrt{T/2+1}}{4e^2\phi(T/2+1)} - 2\phi(t_1)\sqrt{t_1} \geq \frac{\sqrt{T/2+1}}{8e^2\phi(T/2+1)} ,$$

which plugged back into Eq. (5) gives

$$\begin{aligned}\sum_{k=1}^{T/2} k\eta_k^2 &\geq \frac{\frac{T}{2^9 e^4 \phi(T/2+1)^4}}{\left(\frac{1}{2} - \frac{1}{2^9 e^4 \phi(T/2+1)^4}\right) T} \cdot \frac{T/2}{8e^2\phi(T/2+1)^4} \\ &\geq \frac{T}{2^{13} e^8 \phi(T+1)^8} .\end{aligned}$$

Propagated back into Eq. (4), we get that

$$128\phi(T+1)^4 \geq \frac{H_{T/2}-1}{T} \cdot \frac{T}{2^{13}e^8\phi(T+1)^8} ,$$

or equivalently,

$$\phi(T+1)^{12} \geq (H_{T/2}-1) \cdot \frac{1}{2^{20}e^8} .$$

Lastly, by recalling that the harmonic sum grows logarithmically, we get

$$\phi(T+1) \geq \frac{(H_{T/2}-1)^{1/12}}{2^{5/3}e^{2/3}} \geq \frac{1}{2^{5/3}e^{2/3}} \cdot (\log^{1/12}(T/2) - 1) = \Omega(\log^{1/12}(T)) ,$$

which completes the proof.

6 Discussion

In this work, we proved that the last iterate of GD cannot converge at the information-theoretically optimal rate whenever the stopping time is not chosen in advance. An alternative phrasing of this result is that in order to achieve optimal convergence rate with GD in the classic Lipschitz convex setting, either a priori knowledge of the time horizon, or else averaging, is required. As discussed, this proves the conjecture of Jain et al. [2019] for SGD, which thus far remained open, and even extends the result to GD.

Our work leaves open several follow-up questions:

- First, it would be of interest to address the strongly-convex case, where there exist similar gaps between the anytime $\log T/T$ last iterate rate [Shamir and Zhang, 2013] as opposed to $1/T$ when T is chosen in advance [Jain et al., 2019]. We believe our technique should be useful in order to resolve this setting as well, although the technical details still remain unclear at the moment.
- Another open direction is given by noting that while our result is qualitatively stronger than anticipated as it applies even to deterministic GD, it is quantitatively weaker than the known anytime upper bound (Example 2), leaving open a fine-grained gap between $\log^{1/12}(T)/\sqrt{T}$ and $\log(T)/\sqrt{T}$.
- Finally, we note that a stepsize sequence can have an optimally converging *sub*-sequence with $(t_k)_{k=0}^{\infty} \subsetneq \mathbb{N}$ satisfying $\mathcal{E}_{\boldsymbol{\eta}}^{G,D}(t_k) = O(DG/\sqrt{t_k})$. For example, applying a "doubling trick" by concatenating optimal time-dependent stepsize sequences (e.g., as provided by Jain et al. 2019) for increasing powers of 2, yields an infinite sequence $\boldsymbol{\eta}$ so that $\mathcal{E}_{\boldsymbol{\eta}}^{G,D}(2^k) = O(DG/\sqrt{2^k/2}) = O(DG/\sqrt{2^k})$. However, this leads to an exponentially sparse subset of optimal stopping times, and in particular the set of optimal stopping times has density zero.[3] It is therefore interesting to ask how "dense" suboptimal stopping times can be, and whether the sparsity exhibited by the doubling trick is inevitable: Is it true that for any stepsize sequence, any sub-sequence of stopping times which converge optimally necessarily has density zero? Notably, proving this claim would strengthen our result from showing that the set of suboptimal stopping times is nonempty, to showing that it has positive density.

[3] A set $\mathcal{T} \subset \mathbb{N}$ is called a zero-density set if $\lim_{T\to\infty} \frac{|\mathcal{T}\cap\{1,\ldots,T\}|}{T} = 0$.

References

Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, pages 1–14, 2024.

Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. *arXiv preprint arXiv:2507.11274*, 2025.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics*, pages 177–186. Springer, 2010.

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. In *International Conference on Learning Representations*, 2025.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *Advances in Neural Information Processing Systems*, 37: 9974–10007, 2024.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. From continual learning to sgd and back: Better rates for continual linear models. In *Fourth Conference on Lifelong Learning Agents-Workshop Track*, 2025.

Guillaume Garrigos, Daniel Cortild, Lucas Ketels, and Juan Peypouquet. Last-iterate complexity of sgd for convex and smooth stochastic problems. *arXiv preprint arXiv:2507.14122*, 2025.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

Benjamin Grimmer, Kevin Shu, and Alex L Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *INFORMS Journal on Optimization*, 7(2):156–169, 2025.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019.

Guy Kornowski and Ohad Shamir. Open problem: Anytime convergence rate of gradient descent. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5335–5339. PMLR, 2024.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. *arXiv preprint arXiv:2403.07723*, 2024a.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024b.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Arkadi Semenovich Nemirovski and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley, 1983.

Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1571–1578, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Ohad Shamir. Open problem: Is averaging needed for strongly convex stochastic gradient descent? In *Conference on Learning Theory*, pages 47–1. JMLR Workshop and Conference Proceedings, 2012.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pages 2934–2992. PMLR, 2019.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

Zihan Zhang, Jason Lee, Simon Du, and Yuxin Chen. Anytime acceleration of gradient descent. In *Proceedings of Thirty Eighth Conference on Learning Theory*, pages 5991–6013. PMLR, 2025.

A Proof of Lemma 8

To prove Lemma 8, we start by proving the following auxiliary result.

Lemma 9. *Let $T \in \mathbb{N}$, and suppose $(a_j)_{j=0}^T, (b_j)_{j=0}^T \geq 0$ are non-negative sequences satisfying the following conditions:*

(i) $\sum_{j=0}^T a_j^2 \leq \frac{1}{2}$.

(ii) For all j it holds that $b_j \leq \min\{\frac{1}{2}, \frac{1}{2\eta_j\sqrt{T+1}}\}$.

(iii) For all j it holds that $a_j \sum_{k=j+1}^T \eta_k \leq \frac{1}{2}\eta_j b_j$.

Then $\mathcal{E}_{\boldsymbol{\eta}}(T) \geq \frac{1}{2}\sum_{j=0}^{T-1} a_j b_j \eta_j$.

Proof of Lemma 9. For $i \in \{0, \ldots, T\}$ we define the vectors

$$\mathbf{v}_i := \sum_{j=1}^{i} a_{j-1}\mathbf{e}_j - b_i \mathbf{e}_{i+1} \in \mathbb{R}^{T+1} ,$$

and let $f : \mathbb{R}^{T+1} \to \mathbb{R}$ be the function

$$f(\mathbf{x}) := \max_i \{\mathbf{v}_i^\top \mathbf{x}\} .$$

Note that each component $\mathbf{x} \mapsto \mathbf{v}_i^\top \mathbf{x}$ is 1-Lipschitz since $\|\mathbf{v}_i\|^2 \leq \sum_{j=0}^{T-1} a_j^2 + \max_j b_j^2 \leq 1$, and therefore f is 1-Lipschitz as the maximum of 1-Lipschitz functions. We consider GD applied to f, initialized at the origin and projected onto the unit ball $\mathbb{B} \subset \mathbb{R}^{T+1}$, where the given subgradient at a point $\mathbf{x} \in \mathbb{R}^{T+1}$ corresponds to $\mathbf{v}_i$ for the minimal index i such that $f(\mathbf{x}) = \mathbf{v}_i^\top \mathbf{x}$.

We will first prove by induction over t that for all $t \in [T]$:

$$\mathbf{x}_t = \sum_{j=1}^{t} \Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t-1}\eta_k\Big)\mathbf{e}_j . \tag{6}$$

For the base case $t = 1$, note that all the $\mathbf{v}_i$'s are in the sub-gradient set of $\mathbf{x}_0 = \mathbf{0}$, so the sub-gradient choice of the minimal index leads to

$$\mathbf{x}_1 = \mathrm{Proj}_{\mathbb{B}}(\mathbf{0} - \eta_0 \mathbf{v}_0) = \mathrm{Proj}_{\mathbb{B}}(\eta_0 b_0 \mathbf{e}_1) \overset{(\eta_0 b_0 \leq \frac{1}{2\sqrt{T+1}} < 1)}{=} \eta_0 b_0 \mathbf{e}_1 ,$$

and therefore (6) holds for $t = 1$. Now assume (6) holds at time t. To obtain the claim for $t+1$, we start by showing that the returned sub-gradient at $\mathbf{x}_t$ is $\mathbf{v}_t$. To see that, we see that for any $i \in \{0, \ldots, T-1\}$:

$$\begin{aligned}
\mathbf{v}_i^\top \mathbf{x}_t &= \left(\sum_{j=1}^{i} a_{j-1}\mathbf{e}_j - b_i \mathbf{e}_{i+1}\right)^\top \left(\sum_{j=1}^{t} \Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t-1}\eta_k\Big)\mathbf{e}_j\right) \\
&= \sum_{j=1}^{\min\{i,t\}} a_{j-1} \underbrace{\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t-1}\eta_k\Big)}_{(\star)} - \mathbb{1}_{\{i+1\leq t\}} \cdot b_i \underbrace{\Big(b_i\eta_i - a_i\textstyle\sum_{k=i+1}^{t-1}\eta_k\Big)}_{(\star\star)} ,
\end{aligned}$$

and note that $(\star), (\star\star) > 0$ by assumption *(iii)*, and therefore the minimal index that maximizes the expression above is clearly $i = t$. Hence, the given sub-gradient at $\mathbf{x}_t$ is $\mathbf{v}_t$ as claimed, which gives

that

$$\begin{aligned}
\mathbf{x}_{t+1} &= \mathrm{Proj}_{\mathbb{B}}(\mathbf{x}_t - \eta_t \mathbf{v}_t) \\
&= \mathrm{Proj}_{\mathbb{B}}\left(\sum_{j=1}^{t}\Big(b_{j-1}\eta_{j-1} - a_{j-1}\sum_{k=j}^{t-1}\eta_k\Big)\mathbf{e}_j - \eta_t\sum_{j=1}^{t}a_{j-1}\mathbf{e}_j + \eta_t b_t \mathbf{e}_{t+1}\right) \\
&= \mathrm{Proj}_{\mathbb{B}}\left(\sum_{j=1}^{t+1}\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t}\eta_k\Big)\mathbf{e}_j\right) \\
&= \sum_{j=1}^{t+1}\underbrace{\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t}\eta_k\Big)}_{(\diamond_j)}\mathbf{e}_j \qquad (7)
\end{aligned}$$

where the last equality follows from noting that for each j, by assumption *(iii)* it holds that

$$b_{j-1}\eta_{j-1} \geq (\diamond_j) \geq b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{T}\eta_k \geq \frac{1}{2}b_{j-1}\eta_{j-1} \geq 0\,,$$

hence

$$\sum_{j=1}^{t+1}(\diamond_j)^2 \leq \sum_{j=1}^{T+1}(b_{j-1}\eta_{j-1})^2 \leq \sum_{j=1}^{T+1}\frac{1}{4(T+1)} < 1$$

using assumption *(ii)*. We see that (7) completes the induction step, proving (6).

To complete the proof, we note that $\min_{\mathbf{x}\in\mathbb{B}} f(\mathbf{x}) \leq f(\mathbf{0}) = 0$, and recall we saw that $f(\mathbf{x}_T) = \mathbf{v}_T^\top \mathbf{x}_T$, therefore

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\eta}}(T) &\geq f(\mathbf{x}_T) - \min_{\mathbf{x}\in\mathbb{B}} f(\mathbf{x}) \\
&\geq f(\mathbf{x}_T) \\
&= \mathbf{v}_T^\top \mathbf{x}_T \\
&= \left(\sum_{j=1}^{T} a_{j-1}\mathbf{e}_j - b_T\mathbf{e}_{T+1}\right)^\top \left(\sum_{j=1}^{T}\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{T-1}\eta_k\Big)\mathbf{e}_j\right) \\
&= \sum_{j=1}^{T} a_{j-1}\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{T-1}\eta_k\Big) \\
&\geq_{(iii)} \frac{1}{2}\sum_{j=1}^{T} a_{j-1}b_{j-1}\eta_{j-1} \\
&= \frac{1}{2}\sum_{j=0}^{T-1} a_j b_j \eta_j\,.
\end{aligned}$$

□

We turn back prove Lemma 8. Given $t \in \mathbb{N}$ and any $j < t$, let

$$\begin{aligned}
a_j &:= \frac{\min\{1, \eta_j\sqrt{t+1}\}}{16\phi(t+1)(t+1-j)}\,, \\
b_j &:= \min\left\{\frac{1}{2}, \frac{1}{2\eta_j\sqrt{t+1}}\right\}\,.
\end{aligned}$$

We note that the conditions in Lemma 9 hold:

(i) It holds that

$$\begin{aligned}\sum_{j=0}^{t} a_j^2 &\leq \frac{1}{256\phi(t+1)} \sum_{j=0}^{t} \frac{1}{(t+1-j)^2} \\ &= \frac{1}{256\phi(t+1)} \sum_{j=1}^{t} \frac{1}{j^2} \\ &< \frac{1}{256} \cdot \frac{\pi^2}{6} \\ &< \frac{1}{2} \,.\end{aligned}$$

(ii) For all j :

$$b_j \leq \min\left\{\frac{1}{2}, \frac{1}{2\eta_j\sqrt{t+1}}\right\} \,.$$

(iii) For all j :

$$\begin{aligned}a_j \sum_{k=j+1}^{t} \eta_k &\overset{\text{Lemma 7}}{\leq} \frac{\min\{1, \eta_j\sqrt{t+1}\}}{16\phi(t+1)(t+1-j)} \cdot 2\phi(t+1)(\sqrt{t}-\sqrt{j}) \\ &= \frac{(\sqrt{t}-\sqrt{j})\sqrt{t+1}}{t+1-j} \cdot \min\left\{\frac{1}{8\sqrt{t+1}}, \frac{\eta_j}{8}\right\} \\ &\leq \underbrace{\frac{t+1-\sqrt{j}\sqrt{t+1}}{t+1-j}}_{\leq 1} \cdot \min\left\{\frac{1}{8\sqrt{t+1}}, \frac{\eta_j}{8}\right\} \\ &\leq \min\left\{\frac{1}{8\sqrt{t+1}}, \frac{\eta_j}{8}\right\} \\ &< \frac{1}{2}\eta_j b_j \,.\end{aligned}$$

Therefore, as the conditions of Lemma 9 hold, we can apply it and get that

$$\begin{aligned}\mathcal{E}_{\boldsymbol{\eta}}(t) &\geq \frac{1}{2}\sum_{j=0}^{t-1} a_j b_j \eta_j \\ &= \sum_{j=0}^{t-1} \frac{\min\{1, \eta_j\sqrt{t+1}\}}{16\phi(t+1)(t+1-j)} \cdot \min\left\{\frac{1}{4\sqrt{t+1}}, \frac{\eta_j}{4}\right\} \\ &= \frac{1}{64\phi(t+1)\sqrt{t+1}} \sum_{j=0}^{t-1} \frac{\min\{1, \eta_j\sqrt{t+1}\}^2}{(t+1-j)} \,.\end{aligned}$$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the concrete contributions of this paper, as indicated throughout the rest of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The discussion section explicitly discusses questions left open by our work. Other than that, any theoretical paper is limited by the setting it considers, and we do not believe our paper is further missing pointers to any specific limitations.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions, theorems, lemmas and proofs are provided either in the main text or else in the appendix

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics, and the research conforms with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is theoretical, and its goal is to advance our understanding of the theoretically-optimal last iterate convergence of SGD and GD. While there are potential societal consequences of Machine Learning as a whole, we believe none of which must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.