OPT2025: 17th Annual Workshop on Optimization for Machine Learning

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

Guy Kornowski GUY.KORNOWSKI@WEIZMANN.AC.IL
Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

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. Our proof further suggests that such (slightly) suboptimal stopping times are unavoidably common.

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 $\mathbb{E}[\mathbf{g}_t] \in \partial f(\mathbf{x}_t)$, 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 rate by an excess log factor. 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 1.1.

Ultimately, Jain et al. [2019] proved that the last iterate of SGD 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-

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

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. In many applications however, it is desirable to have anytime algorithms with anytime guarantees, and the anytime nature of SGD is arguably one of its appeals in the first place. 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 final iterate of SGD."*

The conjecture so far remained unproven, leaving 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 and answer 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, we present in Section 2 the formal setting that we consider. In Section 3 we state our main result, and discuss some notable consequences. In Section 4 we present the key proof ideas, with the formal proofs deferred to the appendix. We conclude in Section 5.

1.1. 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.

2. Preliminaries

Notation. We denote $\mathbb{N} := \{1, 2, \dots\}$ 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 $A \subset \mathbb{R}^d$, $\mathbf{x} \in \mathbb{R}^d$. $\partial f(\cdot)$ 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 $\mathbf{x}, \mathbf{y} \in \mathcal{X}$: $|f(\mathbf{x}) - f(\mathbf{y})| \le G\|\mathbf{x} - \mathbf{y}\|$. 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 $\boldsymbol{\eta} = (\eta_t)_{t=0}^{\infty}$.

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 $\mathcal{E}_{\boldsymbol{\eta}}^{G,D} : \mathbb{N} \to \mathbb{R}_{\ge 0}$ is an anytime convergence rate guarantee for $\boldsymbol{\eta}$, if for all G-Lipschitz and convex f over domain $\mathcal{X}$ with diameter at most D, and initialization $\mathbf{x}_0 \in \mathcal{X}$, it holds that*

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

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

3. Main Result

We are now ready to present our main result:

Theorem 3 *No stepsize sequence $\boldsymbol{\eta}$ has an anytime convergence guarantee satisfying*

$$\mathcal{E}_{\boldsymbol{\eta}}^{G,D}(t) = o\left(\frac{DG\log^{1/8}(t)}{\sqrt{t}}\right) \;\; \textit{as } t \to \infty\,.$$

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

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 not knowing when the algorithm should stop. We discuss this in more detail in Section 4.*

Remark 5 (suboptimal stopping times are common) *Note that a stepsize sequence can have a guarantee $\mathcal{E}_{\boldsymbol{\eta}}^{G,D}$ with a subsequence $(t_k)_{k=0}^{\infty} \subsetneq \mathbb{N}$ 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 $\mathcal{E}_{\boldsymbol{\eta}}^{G,D}(2^k) = O(DG/\sqrt{2^k/2}) = O(DG/\sqrt{2^k})$. It is therefore interesting to ask how "dense" suboptimal stopping times are. Our proof actually shows that for any stepsize sequence, as $T \to \infty$, a uniformly random stopping time $t \in [T]$ suffers from the aforementioned poly-log overhead. In other words, (slightly) suboptimal stopping times necessarily have so-called positive natural density [Niven, 1951], and therefore occur often, in a suitable sense.*

4. Proof Idea

By a rescaling argument, it suffices to prove Theorem 3 for $G = 1$, $D = 2$, and we abbreviate notation accordingly by denoting $\mathcal{E}_{\boldsymbol{\eta}} = \mathcal{E}_{\boldsymbol{\eta}}^{1,2}$. Let $\boldsymbol{\eta}$ be some stepsize sequence, and suppose it has a corresponding anytime guarantee satisfying

$$\mathcal{E}_{\boldsymbol{\eta}}(t) \leq \frac{\phi(t)}{\sqrt{t}} \tag{2}$$

for some non-decreasing function $\phi : \mathbb{N} \to [1, \infty)$ (e.g. $\phi(t) = c_1 \log^{c_2}(t) + c_3$ for some constants c_1, c_2, c_3). Our goal is to show that it cannot be that $\phi(t) = o(\log^{1/8}(t))$.

To that end, we start by establishing two 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 $(\eta_0, \ldots, \eta_{t-1})$.

Lemma 6 *For any stepsize sequence $\boldsymbol{\eta}$ and $t \in \mathbb{N}$ it holds that:*

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

Both of the bounds above holds already in one dimension. The first follows from considering a "v-shaped" function with a minimum at 0, and noting that if x_{t-1} is very close to the minimum, then $|x_t| \approx \eta_{t-1}$, which therefore serves a lower bound on the last-iterate error.[2] Intuitively then, the lower bound on the last-iterate holds due to the algorithm not knowing when it should stop. The second bound follows from considering a quadratic, and it shows that the stepsize sum needs to grow at a rate of roughly $\mathcal{E}_{\boldsymbol{\eta}}(t)^{-1} \geq \sqrt{t}/\phi(t)$.

Next, we establish a third lower bound on $\mathcal{E}_{\boldsymbol{\eta}}(t)$.

2. An easy, somewhat hacky way to prove this bound is simply by considering $f(x) = |x|$ initiated at zero, so that all sub-gradients until time t are $0 \in \partial f(0)$, and then $g_t = 1 \in \partial f(0)$. The proof provided in the appendix however does not rely on an inconsistent sub-gradient choice.

Lemma 7 *For any stepsize sequence $\boldsymbol{\eta}$, any $t \in \mathbb{N}$ and ϕ as in Eq. (2), it holds that*

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

The proof of Lemma 7 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 builds on a lower bound for max-of-linear functions due to Nemirovski and Yudin [1983].

We further note that $\min\{1, \eta_j\sqrt{j+1}\} \geq \eta_j\sqrt{j+1}/\phi(j+1)$, since by Lemma 6.1 and Eq. (2) it holds that $\phi(j+1) \geq \eta_j\sqrt{j+1}$ and $\phi(j+1) \geq 1$. Hence, Eq. (3) 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

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

With a bound on ϕ in hand, it remains to lower bound the sum above, which as discussed in Remark 5, we can obtain even for a constant fraction of $t \in [T]$ as $T \to \infty$. We get this by averaging Eq. (4) over $t \in [T]$, and deriving lower bounds on $\sum_{j=0}^{t-1} \eta_j^2(j+1)$ using Lemma 6 with an ℓ_1/ℓ_2 comparison inequality. The full details appear in the appendix.

5. 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. As discussed, this proves a conjecture of Jain et al. [2019] regarding SGD.

Our work leaves open several follow-up questions. In future work, we plan to extend our techniques to handle 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].

Another open direction is 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 $\log^{1/8} T/\sqrt{T}$ and $\log T/\sqrt{T}$.

Finally, following the discussion in Remark 5, we recall that the doubling trick is the only procedure we are aware of that constructs a subsequence of iterates converging at the optimal rate. It is interesting to note that this leads to exponentially increasing optimal stopping times, and therefore the set of optimal stopping times form a zero-density set.[3] Hence, it is natural to ask whether this is inevitable: Is it true that for any stepsize sequence, any subsequence of stopping times which converge optimally necessarily has density zero? Notably, this corresponds to strengthening our result from suboptimal stopping times having positive density to always having density 1.

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

Acknowledgments. GK is supported by an Azrieli Foundation graduate fellowship.

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.

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.

Ivan Niven. The asymptotic density of sequences. *Bulletin of the American Mathematical Society*, 57(6):420–434, 1951.

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.

Appendix A. Proofs

A.1. Proof of Lemma 6

For the first item, 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 $[-1, 1]$:

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

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 $x_{t-1} = \epsilon - c_\epsilon \sum_{j=0}^{t-2} \eta_j = 0$. If at time $t-1$ the subgradient is given by $-1 \in \partial f(x_{t-1})$, this leads to $x_t = \eta_{t-1}$. Noting that $\eta_{t-1} > \epsilon$, we get that $\mathcal{E}_{\boldsymbol{\eta}}(t) \ge f(x_t) = \eta_{t-1} - \epsilon + c_\epsilon \epsilon$. Since this inequality holds for arbitrarily small ϵ, it must hold that

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

We now turn to prove the second claim, which is essentially the proof of [Kornowski and Shamir, 2024, Theorem 1] restated here for completeness. Given $\boldsymbol{\eta}$ and t, consider the convex quadratic

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

Note that f is 1-Lipschitz over $[-1, 1]$ as long as $\sum_{j=0}^{t-1} \eta_j \ge 1$, and is minimized at $f(0) = 0$. Examining the iterates of GD initialized at $x_0 = 1$, a simple induction reveals that $x_t = \prod_{j=0}^{t-1} (1 - \frac{\eta_j}{\sum_{j=0}^{t-1} \eta_j})$. Therefore

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

A.2. Proof of Lemma 7

To prove Lemma 7, we start by proving the following auxiliary result.

Lemma 8 *Let $T \in \mathbb{N}$, and suppose $(a_j)_{j=0}^T, (b_j)_{j=0}^T \ge 0$ 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 [Proof of Lemma 8] 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 $t \in [T]$:

$$\mathbf{x}_t = \sum_{j=1}^{t} \Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t-1}\eta_k\Big)\mathbf{e}_j \,. \tag{5}$$

For the base case $t = 1$, note that all the $\mathbf{v}_i$'s are in the subgradient set of $\mathbf{x}_0 = \mathbf{0}$, so the subgradient 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 (5) holds for $t = 1$. Now assume (5) holds at time t. To obtain the claim for $t+1$, we start by showing that the returned subgradient 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 subgradient at $\mathbf{x}_t$ is $\mathbf{v}_t$ as claimed, which gives

that

$$\begin{aligned}
\mathbf{x}_{t+1} &= \mathrm{Proj}_{\mathbb{B}}(\mathbf{x}_t - \eta_t \mathbf{v}_t) \\
&= \mathrm{Proj}_{\mathbb{B}}\left(\sum_{j=1}^{t}\Big(b_{j-1}\eta_{j-1} - a_{j-1}\sum_{k=j}^{t-1}\eta_k\Big)\mathbf{e}_j - \eta_t\sum_{j=1}^{t}a_{j-1}\mathbf{e}_j + \eta_t b_t \mathbf{e}_{t+1}\right) \\
&= \mathrm{Proj}_{\mathbb{B}}\left(\sum_{j=1}^{t+1}\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t}\eta_k\Big)\mathbf{e}_j\right) \\
&= \sum_{j=1}^{t+1}\underbrace{\Big(b_{j-1}\eta_{j-1} - a_{j-1}\textstyle\sum_{k=j}^{t}\eta_k\Big)}_{(\diamond_j)}\mathbf{e}_j \qquad (6)
\end{aligned}$$

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 (6) completes the induction step, proving (5).

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 will also need the following bound on step sums.

Lemma 9 *For any* $t_1 < t_2 \in \mathbb{N}$: $\sum_{j=t_1}^{t_2}\eta_j \leq 2\phi(t_2+1)(\sqrt{t_2} - \sqrt{t_1})$

Proof [Proof of Lemma 9] By Lemma 6.1, for any $t_1 < t_2 \in \mathbb{N}$:

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

■

We turn back prove Lemma 7. Given $t \in \mathbb{N}$ and any $j < t$, let

$$a_j := \frac{\min\{1, \eta_j\sqrt{t+1}\}}{16\phi(t+1)(t+1-j)}\,, \quad b_j := \min\{\frac{1}{2}, \frac{1}{2\eta_j\sqrt{t+1}}\}\,.$$

We note that the conditions in Lemma 8 hold:

(i) It holds that

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

(ii) For all j :

$$b_j \le \min\{\frac{1}{2}, \frac{1}{2\eta_j\sqrt{t+1}}\}\,.$$

(iii) For all j :

$$\begin{aligned}
a_j \sum_{k=j+1}^{t} \eta_k &\overset{\text{Lemma 9}}{\le} \frac{\min\{1, \eta_j\sqrt{t+1}\}}{16\phi(t+1)(t+1-j)} \cdot 2\phi(t+1)(\sqrt{t} - \sqrt{j+1}) \\
&= \frac{(\sqrt{t} - \sqrt{j+1})\sqrt{t+1}}{t+1-j} \min\{\frac{1}{8\sqrt{t+1}}, \frac{\eta_j}{8}\} \\
&\le \underbrace{\frac{t+1-\sqrt{j+1}\sqrt{t+1}}{t+1-j}}_{\le 1} \cdot \min\{\frac{1}{8\sqrt{t+1}}, \frac{\eta_j}{8}\} \\
&\le \min\{\frac{1}{8\sqrt{t+1}}, \frac{\eta_j}{8}\} < \frac{1}{2}\eta_j b_j\,.
\end{aligned}$$

Therefore, since the conditions of Lemma 8 hold, we can apply and get that

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

A.3. Completing the proof of Theorem 3

We continue the proof following Eq. (4):

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

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

$$\begin{aligned}
128\phi(T+1)^4 \geq \frac{1}{T}\sum_{t=1}^{T} 128\phi(t+1)^4 &\geq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{t-1}\frac{j\eta_j^2}{(t+1-j)} \\
&= \frac{1}{T}\sum_{k=1}^{T-1} k\eta_k^2 \sum_{t=1}^{T}\sum_{j=0}^{t-1}\frac{\mathbb{1}\{j=k\}}{(t+1-j)} \\
&= \frac{1}{T}\sum_{k=1}^{T-1} k\eta_k^2 \sum_{t=k}^{T}\frac{1}{(t+1-k)} \\
&= \frac{1}{T}\sum_{k=1}^{T-1} k\eta_k^2 H_{T+1-k} \qquad \# \text{ where } \; H_n := \sum_{i=1}^{n}\frac{1}{i} \\
&\geq \frac{H_{T/2}}{T}\sum_{k=1}^{T/2} k\eta_k^2 \,. \qquad (7)
\end{aligned}$$

Moreover, for $t_1 < T$ soon to be chosen it holds that

$$\begin{aligned}
\sum_{k=1}^{T/2} k\eta_k^2 \geq \sum_{k=t_1}^{T/2} k\eta_k^2 &\geq t_1 \sum_{k=t_1}^{T/2}\eta_k^2 \\
&\overset{(1)}{\geq} \frac{t_1}{T/2-t_1}\left(\sum_{k=t_1}^{T/2}\eta_k\right)^2 \\
&\overset{(2)}{\geq} \frac{t_1}{T/2-t_1}\left(\frac{\sqrt{T/2+1}}{2e^4\phi(T/2+1)} - 2\phi(t_1)\sqrt{t_1+1}\right)^2 , \qquad (8)
\end{aligned}$$

where (1) follows from a standard ℓ_1/ℓ_2 inequality (for every vector $x \in \mathbb{R}^n : \|x\|_2^2 \geq \|x\|_1^2/n$), and (2) is due to the fact that using Lemma 6, it holds that

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

By setting $t_1 := \lfloor \frac{T/2+1}{64e^8\phi(T/2+1)^2} \rfloor - 1$, it holds that

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

which plugged back into Eq. (8) gives

$$\begin{aligned}\sum_{k=1}^{T/2} k\eta_k^2 &\geq \frac{\frac{T}{256e^8\phi(T/2+1)^2}}{\left(\frac{1}{2}-\frac{1}{256e^8\phi(T/2+1)^2}\right)T}\cdot\frac{T/2}{16e^8\phi(T/2+1)^2}\\ &\geq \frac{T}{2^{13}e^{16}\phi(T+1)^4}\,.\end{aligned}$$

Going back to Eq. (7), we therefore see that

$$128\phi(T+1)^4 \geq \frac{H_{T/2}}{T}\cdot\frac{T}{2^{13}e^{16}\phi(T+1)^4}\,.$$

By rearranging and recalling that the harmonic sum grows logarithmically, we overall get

$$\phi(T+1) \geq \frac{(H_{T/2})^{1/8}}{2^{5/2}e^2} \geq \frac{1}{2^{5/2}e^2}\cdot\log^{1/8}(T/2)\,,$$

which completes the proof.