Open Problem: Anytime Convergence Rate of Gradient Descent

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

Editors: Shipra Agrawal and Aaron Roth

Abstract

Recent results show that vanilla gradient descent can be accelerated for smooth convex objectives, merely by changing the stepsize sequence. We show that this can lead to surprisingly large errors indefinitely, and therefore ask: Is there any stepsize schedule for gradient descent that accelerates the classic $\mathcal{O}(1/T)$ convergence rate, at *any* stopping time T?

1. Introduction

Consider the classic setting of optimizing a smooth convex objective via gradient descent (GD): Given a convex function $f:\mathbb{R}^d\to\mathbb{R}$ which is L-smooth (i.e. ∇f exists and is L-Lipschitz), and an initial point $x_0\in\mathbb{R}^d$, the GD iterates with stepsizes $(\eta_t)_{t=0}^{\infty}$ are defined as $x_{t+1}=x_t-\eta_t\nabla f(x_t)$.

The textbook analysis of GD under this setting (e.g. Nesterov, 2018; Bubeck, 2015) asserts that when the stepsize schedule is fixed to be constant $\eta_t\equiv\overline{\eta}\in(0,\frac{2}{L})$, the iterates satisfy the bound

$$f(x_T)-f^*\lesssim\frac{L\left\|x_0-x^*\right\|^2}{T}\quad\text{for all } T\in\mathbb{N}\,, \tag{1}$$

where $f^*=\inf f$, x^* is any minimizer of f, and "$\lesssim$" hides a constant. It is also well-known that for constant steps larger than $2/L$, the algorithm can diverge. The behavior of GD in this setting is extremely well-studied, and one would think that it is fully understood.

However, quite unexpectedly, a recent line of work established that GD can achieve faster convergence rates than implied by Eq. (1), without any modification to the algorithm itself, merely by using appropriate *non-constant* stepsize schedules which incorporate occasional *long steps*, larger than $2/L$ (Grimmer et al., 2023; Altschuler and Parrilo, 2023a).[1] In particular, Altschuler and Parrilo (2023b) constructed a stepsize sequence, coined the "silver stepsize" schedule, that guarantees

$$f(x_T)-f^*\lesssim\frac{L\left\|x_0-x^*\right\|^2}{T^{\log_2(1+\sqrt{2})}}\approx\frac{L\left\|x_0-x^*\right\|^2}{T^{1.2716}}\quad\text{for all } T=2^n-1, n\in\mathbb{N}\,. \tag{2}$$

Remarkably, compared to Eq. (1), this bound achieves an accelerated $o(1/T)$ rate, merely by changing the GD stepsize schedule. However, note that the bound no longer applies for all T. Instead, it only applies for certain exponentially-increasing horizons $T=2^n-1$, with no guarantee on the performance of intermediate iterates. From a practical viewpoint, this is not quite satisfactory, as often the number of iterations is not carefully chosen in advance. Although this can be circumvented by doubling tricks or tracking the best iterate obtained so far (as further discussed below), in practice it is desirable to have a uniform, monotonically-decreasing guarantee on the error, which ensures that at any large enough stopping point, the resulting optimization error is small. As far as we know, none of the existing results for stepsize-based acceleration apply in an anytime fashion, and it is not clear that such acceleration is even possible. Hence, we formulate the following open problem:

1. (Daccache, 2019; Eloi, 2022; Das Gupta et al., 2024; Grimmer, 2023) previously improved the *constant* factor in (1).

Open Problem 1 *What is the best **anytime** convergence rate achievable by GD with some stepsize sequence $(\eta_t)_{t=0}^{\infty}$, uniformly over L-smooth convex functions? In particular, is there any stepsize sequence and some $\alpha > 1$, such that for all L-smooth convex f,*

$$f(x_T) - f^* \lesssim \frac{L\left\|x_0 - x^*\right\|^2}{T^{\alpha}} \quad \textit{for all } T \in \mathbb{N}\,? \tag{3}$$

We remark that we seek an anytime, monotonically decreasing *upper bound* on the error: Indeed, with long steps, the errors themselves may not decrease monotonically. A similar phenomenon is exhibited by Nesterov's accelerated gradient method (Nesterov, 1983): It is well-known (e.g., d'Aspremont et al., 2021) that this algorithm does *not* monotonically decrease the error, while still having an anytime, monotonically decreasing error bound similar to Eq. (1) (replacing T^{-1} by T^{-2}, which is the optimal dimension-free rate for gradient-based algorithms).

We further note that while we focus on the convex setting, the analogous question for the *strongly*-convex case is also of interest, for which it is unclear whether any stepsize sequence achieves an anytime $\exp(-T/o(\kappa))$ rate, uniformly over functions with condition number κ.

Equivalent view: Bounds on the iterate vs. best iterate A bound such as (2) can be easily converted to an anytime bound on $\min_{t\in[T]} f(x_t) - f^*$ (namely, the best iterate obtained so far): Indeed, for any given $T \in \mathbb{N}$, let $\hat{T}$ be the largest integer such that $\hat{T} \leq T$ and $\hat{T} = 2^n - 1$ for some $n \in \mathbb{N}$. It is easy to show that $\hat{T} \geq T/2$, and hence by (2), $\min_{t\in[T]} f(x_t) - f^* \;\leq\; f(x_{\hat{T}}) - f^* \lesssim \frac{L\|x_0-x^*\|^2}{\hat{T}^{1.2716}} \leq \frac{2^{1.2716}L\|x_0-x^*\|^2}{T^{1.2716}}$ for all $T \in \mathbb{N}$. This is an anytime guarantee, which matches (2) up to a small numerical constant (indeed, Grimmer et al. (2023) even present their accelerated result in this manner). However, this anytime guarantee no longer applies to individual iterates x_T. Thus, our question is equivalent to asking whether an appropriate stepsize schedule can accelerate GD in terms of $f(x_T) - f^*$, rather than $\min\limits_{t\in[T]} f(x_t) - f^*$.

2. Preliminary results

We take steps towards the resolution of Open Problem 1 by providing two results, both of which hold already in dimension $d = 1$. These results indicate the tension between acceleration with GD and anytime guarantees. Moreover, we establish that current accelerating stepsize schedules can strongly fail to meet anytime guarantees.

First, we note that long steps are not only required for anytime acceleration, in fact such acceleration necessitates *arbitrarily large* steps.[2] This can be formally stated as follows:

Theorem 1 *Suppose that GD with stepsizes $(\eta_t)_{t=0}^{\infty}$ satisfies an accelerated uniform guarantee, namely $\forall T \in \mathbb{N} : f(x_T) - f^* \lesssim o(L/T)$ for any L-smooth convex f. Then $\limsup_{t\to\infty}\eta_t = \infty$.*

Indeed, the step size schedules in Grimmer et al. (2023) and Altschuler and Parrilo (2023b) satisfy this requirement: They both involve a fractal-like stepsize schedule, where the stepsize value increases exponentially at exponentially-increasing intervals. However, our next result implies that occasional huge steps (compared to previous steps) can prevent any decaying uniform bound, even one which is not accelerated. As a corollary (see Corollary 4), this implies that the silver stepsize schedule cannot enjoy *any* convergence guarantee which holds in an anytime fashion.

2. Note that this assertion relies on the finite-time bound being uniform over all smooth convex functions. In fact, an asymptotic $o(L/T)$ for any *fixed* function actually holds for constant stepsizes (Lee and Wright, 2019).

Theorem 2 *For any $L > 0$, stepsize schedule $(\eta_t)_{t=0}^{\infty}$ and $T \in \mathbb{N}$ satisfying $\min\{\frac{\eta_T}{2}, \sum_{t=0}^{T-1}\eta_t\} \geq \frac{1}{L}$, there exists an L-smooth convex $f : \mathbb{R} \to \mathbb{R}$ such that $f(x_{T+1}) - f^* \geq \frac{1}{32}L\|x_0 - x^*\|^2 \left(\frac{\eta_T}{\sum_{t=0}^{T-1}\eta_t}\right)^2$.*

Note that the lower bound holds for T as long as it satisfies $\eta_T \geq \frac{2}{L}$ and $\sum_{t=0}^{T-1}\eta_t \geq \frac{1}{L}$. The latter condition is in a sense generic, and should be expected to hold for all large enough T, since otherwise GD cannot guarantee convergence to possibly far-away minima in the first place. Thus the important condition is that $\eta_T \geq \frac{2}{L}$, namely that at time T GD takes a long step (beyond the $2/L$ regime). The theorem formally shows that long steps may "overshoot", as measured by the squared ratio $(\frac{\eta_T}{\sum_{t=0}^{T-1}\eta_t})^2$. In particular, *the larger this ratio is, the larger the error can be after the long step.*

Corollary 3 *If a stepsize sequence $(\eta_t)_{t=0}^{\infty}$ satisfies an anytime accelerated bound as in Eq. (3), then $\eta_T \lesssim \frac{\sum_{t=0}^{T-1}\eta_t}{T^{\alpha/2}} = o\left(\frac{\sum_{t=0}^{T-1}\eta_t}{\sqrt{T}}\right)$ for infinitely many $T \in \mathbb{N}$ in which long steps occur.*

Overall, we see that acceleration requires the stepsize sequence to have a subsequence going to infinity by Theorem 1, yet "not too fast", as captured by Corollary 3. Furthermore, Theorem 2 shows that if a stepsize schedule that incorporates long steps satisfies $\left(\frac{\eta_T}{\sum_{t=0}^{T-1}\eta_t}\right)^2 \gtrsim 1$ for infinitely many $T \in \mathbb{N}$, then the lower bound $f(x_{T+1}) - f^* \gtrsim L\|x_0 - x^*\|^2$ applies for arbitrarily large $T \in \mathbb{N}$.[3] In particular, it is easy to verify that the silver stepsize satisfies this property (Altschuler and Parrilo, 2023b, Eq. 1.3 and Lemma 2.3), hence we get:

Corollary 4 *No anytime bound of the silver stepsize schedule goes to zero (at any rate whatsoever).*

3. Proofs

3.1. Proof of Theorem 1

We first note that by rescaling, it suffices to prove the claim for $L = 1$. We will also assume $\sum_{t=0}^{\infty}\eta_t = \infty$ (otherwise it is well-known that GD may not converge). Next, consider the convex quadratic $f_T(x) = \frac{x^2}{2\sum_{t=0}^{T-1}\eta_t}$ which is minimized at $f_T(0) = 0$, and note that for sufficiently large T we may assume without loss of generality that f_T is 1-smooth, and $\forall t < T : \eta_t \leq \frac{1}{2}\sum_{j=0}^{T-1}\eta_j$.[4] For $x_0 = 1$, a simple induction reveals that $x_T = \prod_{t=0}^{T-1}\left(1 - \frac{\eta_t}{\sum_{t=0}^{T-1}\eta_t}\right)$. So if for all $T \in \mathbb{N}$: $f(x_T) - f^* = f(x_T) \leq \phi(T)$ for some $\phi(T) = o(1/T)$, then

$$\begin{aligned}
\phi(T) \geq f_T(x_T) &= \frac{1}{2\sum_{t=0}^{T-1}\eta_t} \cdot \prod_{t=0}^{T-1}\left(1 - \frac{\eta_t}{\sum_{t=0}^{T-1}\eta_t}\right)^2 \\
&= \frac{1}{2\sum_{t=0}^{T-1}\eta_t} \cdot \exp\left[2 \cdot \sum_{t=0}^{T-1}\log\left(1 - \frac{\eta_t}{\sum_{t=0}^{T-1}\eta_t}\right)\right] \\
&\geq \frac{1}{2\sum_{t=0}^{T-1}\eta_t} \cdot \exp\left[-4 \cdot \sum_{t=0}^{T-1}\frac{\eta_t}{\sum_{t=0}^{T-1}\eta_t}\right] = \frac{e^{-4}}{2\sum_{t=0}^{T-1}\eta_t},
\end{aligned}$$

thus $\max_{0 \leq t \leq T-1}\eta_t \geq \frac{1}{T}\sum_{t=0}^{T-1}\eta_t \geq \frac{e^{-4}}{2T\phi(T)} \xrightarrow{T\to\infty} \infty$.

3. This is the strongest possible lower bound, since $f(x_0) - f^* \leq \frac{L}{2}\|x_0 - x^*\|^2$ in the first place due to smoothness.
4. Otherwise, since $\sum_{t=0}^{\infty}\eta_t = \infty$, there is a subsequence of stepsizes diverging to ∞, proving the theorem altogether.

3.2. Proof of Theorem 2

We first note that by rescaling, it suffices to prove the claim for $L = 1$. Accordingly, we let $T \in \mathbb{N}$ be so that $\min\{\frac{\eta_T}{2}, \sum_{t=0}^{T-1}\eta_t\} \geq 1$, and denote $\Sigma_T := \sum_{t=0}^{T-1}\eta_t,\ c_T := \frac{\eta_T^2}{32\Sigma_T^2}$.

Let $a, r > 0$ to be determined later, and consider the scaled Huber loss and initialization point:

$$f(x) = \begin{cases} \frac{a}{2}x^2\ , & x \leq r \\ ar\cdot x - \frac{ar^2}{2}\ , & x > r \end{cases}\ , \qquad x_0 = r + ar\Sigma_T\ .$$

Note that f is convex, a-smooth, and that $f^* = f(0) = 0$. We will show that for a suitable choice of $a \leq 1,\ r > 0$:

$$f(x_{T+1}) - f^* \geq c_T\,\|x_0 - x^*\|^2 = \frac{1}{32}\,\|x_0 - x^*\|^2\left(\frac{\eta_T}{\sum_{t=0}^{T-1}\eta_t}\right)^2\ . \tag{4}$$

Lemma 5 *There exist $a, r > 0$ such that* $\max\left\{\frac{2}{\eta_T},\ \left(\frac{8c_T}{\eta_T^2 r^2}\right)^{1/3}\right\} \leq a \leq \min\left\{1,\ \frac{1-r}{r\Sigma_T}\right\}$.

Proof By assumption on T that $\eta_T \geq 2$, and the definition of $c_T = \frac{\eta_T^2}{32\Sigma_T^2} \iff \sqrt{8c_T} = \frac{\eta_T}{2\Sigma_T}$ we get $\frac{\sqrt{8c_T}}{\eta_T} \leq \min\left\{\sqrt{\eta_T c_T},\ \frac{1}{2\Sigma_T}\right\}$. Thus, there exists some $r > 0$ such that

$$\frac{\sqrt{8c_T}}{\eta_T} \leq r \leq \min\left\{\sqrt{\eta_T c_T},\ \frac{1}{2\Sigma_T}\right\}\ . \tag{5}$$

Fixing such r, and recalling that $\Sigma_T \geq 1$ by assumption on T, we get that $r \leq \frac{1}{2\Sigma_T} \leq \frac{1}{2}$ which implies $\frac{1}{8} \leq (1-r)^3$, thus $\frac{1}{2\Sigma_T} \leq \frac{4(1-r)^3}{\Sigma_T} = \frac{\eta_T^2(1-r)^3}{8c_T\Sigma_T^3}$. Combining this with Eq. (5), we get the cruder upper bound $\frac{\sqrt{8c_T}}{\eta_T} \leq r \leq \min\left\{\sqrt{\eta_T c_T},\ \frac{\eta_T^2(1-r)^3}{8c_T\Sigma_T^3}\right\}$. Rearranging the latter inequalities, we get that $\frac{2}{\eta_T} \leq \left(\frac{8c_T}{\eta_T^2 r^2}\right)^{1/3} \leq \min\left\{1,\ \frac{1-r}{r\Sigma_T}\right\}$, so in particular $\max\left\{\frac{2}{\eta_T},\ \left(\frac{8c_T}{\eta_T^2 r^2}\right)^{1/3}\right\} \leq \min\left\{1,\ \frac{1-r}{r\Sigma_T}\right\}$. Thus, setting a between these left hand side and right hand side completes the proof. ■

Following Lemma 5, we consider a, r that satisfy the conditions stated therein. That being the case, since $a \leq 1$ we see that f is indeed 1-smooth. Furthermore, we have for all $t \leq T : x_{t+1} = x_t - \eta_t f'(x_t) = x_t - ar\eta_t$, thus $x_T = x_0 - ar\sum_{j=0}^{T-1}\eta_j = r$. This implies, by the gradient descent update and the definition of f, that $x_{T+1} = x_T - \eta_T f'(x_T) = r - ar\eta_T = r(1 - a\eta_T)$. Further noting that by Lemma 5 it holds that $\frac{2}{\eta_T} < a$ which implies $a\eta_T - 1 \geq 1$, we overall get that

$$\begin{aligned} f(x_{T+1}) - f^* = f(x_{T+1}) &= \frac{a}{2}(x_{T+1})^2 = \frac{a(a\eta_T - 1)^2 r^2}{2} \\ &\overset{(\star)}{\geq} \frac{a^3\eta_T^2 r^2}{8} \overset{(\star\star)}{\geq} c_T \overset{(\star\star\star)}{\geq} c_T\,|r + ar\Sigma_T|^2 = c_T\,|x_0 - x^*|^2\ , \end{aligned}$$

where $(\star)$, $(\star\star)$, $(\star\star\star)$ all follow from Lemma 5, thus establishing Eq. (4), completing the proof.

Acknowledgments

This research is supported in part by European Research Council (ERC) grant 754705. GK is supported by an Azrieli Foundation graduate fellowship.

References

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

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

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Antoine Daccache. Performance estimation of the gradient method with fixed arbitrary step sizes. Master's thesis, Université Catholique de Louvain, 2019.

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

Alexandre d'Aspremont, Damien Scieur, Adrien Taylor, et al. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.

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

Benjamin Grimmer. Provably faster gradient descent via long steps. *arXiv preprint arXiv:2307.06324*, 2023.

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

Ching-pei Lee and Stephen Wright. First-order algorithms converge faster than $o(1/k)$ on convex problems. In *International Conference on Machine Learning*, pages 3754–3762. PMLR, 2019.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

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