LAST-ITERATE CONVERGENCE RATE OF NORMALIZED GRADIENT DESCENT UNDER HÖLDER SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Normalized gradient descent is a widely studied adaptive optimization method. Most existing analyses focus on the best iterate or a weighted average of the iterates, whereas practical implementations typically return the last iterate. In this paper, we study the last-iterate convergence of normalized gradient descent for convex, (ν, M_ν)-Hölder-smooth objectives. For a constant stepsize, we establish an upper bound of $\mathcal{O}\big((\log^2(T)/T)^{(1+\nu)/2}\big)$, which contains a logarithmic overhead relative to the known $\mathcal{O}\big(T^{-(1+\nu)/2}\big)$ guarantees for the best and weighted-average iterates. For $\nu = 0$, this overhead is known to be unavoidable. We complement this analysis with numerical results based on the performance estimation problem (PEP), investigating the finite-horizon worst-case behavior in the smooth setting and whether the logarithmic overhead reflects an intrinsic limitation of constant-step normalized gradient descent. We then show that a linearly decreasing stepsize yields a last-iterate guarantee of $\mathcal{O}\big(T^{-(1+\nu)/2}\big)$, matching the order of the best-iterate/weighted-average guarantees without requiring knowledge of ν and M_ν.

1 INTRODUCTION

Neural networks have achieved remarkable empirical success across a wide range of applications, including computer vision (Russakovsky et al., 2015; Radford et al., 2021; Liu et al., 2021) and natural language processing (Devlin et al., 2019; Brown et al., 2020; DeepSeek-AI et al., 2024). Training such networks is challenging: their loss landscapes are often ill-conditioned, with curvature that varies widely across parameters and during training, which makes the performance of optimization methods highly sensitive to the choice of stepsize. Adaptive optimization methods, which scale their updates using observed gradients, are designed to mitigate this sensitivity. Building on AdaGrad (Duchi et al., 2011), Adam and its variant AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) have become the de facto standard for training deep neural networks, and Muon (Jordan et al., 2024; Liu et al., 2025) has recently emerged as a strong alternative.

One of the most basic adaptive optimization methods is normalized gradient descent. Despite its simple form, it is closely related to modern optimizers: without momentum (and with $\epsilon = 0$), Adam reduces to sign descent and Muon to orthogonalized gradient descent, and all three methods are instances of normalized steepest descent with respect to the ℓ_∞, spectral, and Euclidean norms, respectively (Bernstein & Newhouse, 2024). Adam and Muon are among the most widely used optimizers for training large-scale models (Brown et al., 2020; DeepSeek-AI et al., 2024; Liu et al., 2025). Moreover, normalized gradient descent is universal: the same stepsize choice yields appropriate convergence rates across a wide range of problem classes, including smooth (Mishkin et al., 2024), non-smooth (Nesterov, 2018), Hölder-smooth (Grimmer, 2019; Orabona, 2023), and (L_0, L_1)-smooth objectives (Vankov et al., 2025). However, most existing analyses bound the *best* iterate or a *weighted average* of the iterates, whereas practical implementations typically return the *last* iterate. This mismatch has motivated a growing body of work on last-iterate convergence across a wide range of optimization methods (Shamir & Zhang, 2013; Liu & Zhou, 2024b; Zamani & Glineur, 2024; 2025; Defazio et al., 2024a). To the best of our knowledge, however, the last iterate of normalized gradient descent has so far been analyzed only by Zamani & Glineur (2025), and only for convex functions with bounded (sub)gradients, leaving open its behavior on problem classes that admit unbounded gradients.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Summary of the convergence rates for normalized gradient descent assuming that f is convex and (ν, M_ν)-Hölder smooth. Numerical constants are omitted in the rates.

	Reference	Convergence Criteria	ν	Stepsize	Rate
constant stepsize	Grimmer (2019)	$\min_t f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)$	$\nu \in [0,1]$	$\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$	$M_\nu \left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\right)^{1+\nu}$
	Orabona (2023)	$f\left(\frac{1}{\sum_{t=1}^T \frac{1}{\|\nabla f(\boldsymbol{x}_t)\|}} \sum_{t=1}^T \frac{\boldsymbol{x}_t}{\|\nabla f(\boldsymbol{x}_t)\|}\right) - f(\boldsymbol{x}^\star)$	$\nu \in [0,1]$	$\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$	$M_\nu \left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\right)^{1+\nu}$
	Zamani & Glineur (2025)	$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)$	$\nu = 0$	$\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$	$\frac{M_\nu \|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}(1+\log(T))$
	Theorem 1 **(new)**	$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)$	$\nu \in [0,1]$	$\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$	$M_\nu \left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\right)^{1+\nu}(1+\log(T))^{1+\nu}$
linearly decreasing stepsize	Zamani & Glineur (2025)	$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)$	$\nu = 0$	$\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\left(1-\frac{t}{T}\right)$	$\frac{M_\nu \|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$
	Theorem 2 **(new)**	$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)$	$\nu \in [0,1]$	$\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\left(1-\frac{t}{T+1}\right)$	$M_\nu \left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\right)^{1+\nu}$

In this paper, we analyze the last-iterate convergence rate of normalized gradient descent for convex[1] and Hölder-smooth objectives, a general smoothness class that includes both smooth and non-smooth functions as special cases. We first show that, with a constant stepsize, the last iterate converges slower than the best and weighted-average iterates by a factor of order $\log^{1+\nu}(T)$, where T is the number of iterations. This overhead cannot be removed in general: for $\nu = 0$, our bound matches the lower bound of Zamani & Glineur (2025) for the same constant stepsize, and for the smooth case ($\nu = 1$), numerical results based on the performance estimation problem (PEP) (Drori & Teboulle, 2014; Taylor et al., 2017) indicate that the last-iterate rate is indeed slower than $\mathcal{O}(1/T)$. We then show that a linearly decreasing stepsize removes this overhead, yielding a last-iterate rate that matches the best-iterate and weighted-average guarantees. Table 1 compares our results with the existing ones.

Our contributions are summarized as follows.

- **Last-iterate rate with a constant stepsize.** For convex and (ν, M_ν)-Hölder-smooth objectives, we prove that the last iterate of normalized gradient descent with a constant stepsize satisfies $f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) = \mathcal{O}\big(M_\nu(\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2/T)^{(1+\nu)/2}(1+\log T)^{1+\nu}\big)$ (Theorem 1). This rate is slower than the known best-iterate and weighted-average rates by a factor of order $\log^{1+\nu}(T)$. For $\nu = 0$, it recovers the result of Zamani & Glineur (2025), which is tight in this case.
- **Numerical evidence via PEP.** Using the performance estimation problem (PEP), we provide numerical evidence that, in the smooth case ($\nu = 1$), the last-iterate rate of normalized gradient descent with a constant stepsize is slower than $\mathcal{O}(1/T)$. This suggests that the logarithmic overhead is not an artifact of our analysis.
- **Removing the overhead with a decreasing stepsize.** We show that a linearly decreasing stepsize yields the last-iterate rate $\mathcal{O}\big(M_\nu(\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2/T)^{(1+\nu)/2}\big)$ (Theorem 2), matching the best-iterate and weighted-average guarantees. This stepsize does not require knowledge of ν or M_ν, and for $\nu = 0$ the result recovers that of Zamani & Glineur (2025).

Notation. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ denotes the standard inner product and $\|\boldsymbol{x}\| \coloneqq \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}$ the Euclidean norm. T denotes the number of iterations, and $\log$ denotes the natural logarithm. In formal statements (theorems, lemmas, and propositions), where the dependence on all problem parameters is written explicitly, $\mathcal{O}(\cdot)$ hides only absolute numerical constants, i.e., constants independent of T, ν, M_ν, and $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|$. In informal discussions, such as $\mathcal{O}(1/T)$ or $\mathcal{O}(T^{-\frac{1+\nu}{2}})$, we use $\mathcal{O}(\cdot)$ to indicate the dependence on T only, treating the remaining parameters as constants.

2 Related Work

Convergence Analysis of Normalized Gradient Descent. Normalized gradient descent and its variants are among the simplest adaptive optimization methods and have been studied in a variety of settings (Nesterov, 2018; Grimmer, 2019; Cutkosky & Mehta, 2020; Orabona, 2023; Mishkin et al., 2024; Vankov et al., 2025). Here, we focus on the convex, deterministic setting considered in this paper. Most existing analyses concern the best iterate: Nesterov (2018) and Mishkin et al. (2024) cover both smooth and non-smooth objectives, Grimmer (2019) considers Hölder-smooth objectives,

[1]Although many practically important problems are nonconvex, understanding the convex regime remains important; see App. A for a detailed discussion.

and Vankov et al. (2025) considers (L_0, L_1)-smooth objectives. Orabona (2023) analyzes a weighted average of the iterates instead. However, practical implementations typically return the last iterate rather than the best or an averaged one. To the best of our knowledge, Zamani & Glineur (2025) is the only prior work on the last-iterate convergence of normalized gradient descent. For convex functions with bounded subgradients, they showed that, with the constant stepsize $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$, the last iterate converges at the rate $\mathcal{O}(\log T/\sqrt{T})$, which is slower than the $\mathcal{O}(1/\sqrt{T})$ best-iterate rate, and they established a matching lower bound for this stepsize; moreover, they showed that even the optimally tuned constant stepsize only reduces the overhead to $\sqrt{\log T}$, so that a logarithmic loss cannot be avoided with any constant stepsize. They further showed that a linearly decreasing stepsize improves the last-iterate rate to $\mathcal{O}(1/\sqrt{T})$. However, the bounded-subgradient assumption is restrictive: it excludes, for example, smooth objectives such as quadratics, whose gradients are unbounded on $\mathbb{R}^d$. In contrast, we analyze the last iterate under Hölder smoothness, which covers both smooth and non-smooth objectives, and our results recover those of Zamani & Glineur (2025) as the special case $\nu = 0$.

Last-Iterate Convergence of Optimization Methods. Convergence analyses of optimization methods have traditionally bounded the best or an averaged iterate, whereas practical implementations typically return the last iterate. Motivated by this gap between theory and practice, a growing body of work analyzes the last iterate directly, including stochastic gradient descent (Shamir & Zhang, 2013; Liu & Zhou, 2024b), subgradient methods (Zamani & Glineur, 2025) and their variant with the Polyak stepsize (Zamani & Glineur, 2024), shuffling gradient methods (Liu & Zhou, 2024a), momentum methods (Li et al., 2022), and AdaGrad-Norm (Preobrazhenskaia et al., 2026). A recurring finding is that, under standard stepsize choices, the last iterate can converge more slowly than the best or averaged iterate, typically by a logarithmic factor (Shamir & Zhang, 2013; Zamani & Glineur, 2025), and this loss is unavoidable for several standard stepsize choices (Harvey et al., 2019; Kornowski & Shamir, 2026), and that suitably decaying stepsize schedules can close this gap (Jain et al., 2019; Zamani & Glineur, 2025; Defazio et al., 2024a). In this direction, Defazio et al. (2024a) proposed a general framework that converts regret bounds of online learning algorithms into last-iterate guarantees on $f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)$, which in particular justifies linear-decay schedules. However, this framework does not directly apply to normalized gradient descent: its regret bound holds for the normalized gradients $\nabla f(\boldsymbol{x}_t)/\|\nabla f(\boldsymbol{x}_t)\|$ rather than the gradients themselves, and under Hölder smoothness the gradient norms are not bounded away from zero.

3 Preliminaries

In this section, we introduce the problem setting and assumptions, and then summarize the existing guarantees for normalized gradient descent. We consider the unconstrained minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}), \tag{1}$$

and assume that it admits a minimizer $\boldsymbol{x}^\star$. We study normalized gradient descent:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \tilde{\eta}_t \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \tag{2}$$

where $\tilde{\eta}_t > 0$ is a possibly time-varying stepsize. Since $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = \tilde{\eta}_t$, the stepsize $\tilde{\eta}_t$ determines the length of each step regardless of the gradient magnitude, whereas η_t denotes the resulting effective stepsize applied to $\nabla f(\boldsymbol{x}_t)$. If $\nabla f(\boldsymbol{x}_t) = 0$ for some t, then $\boldsymbol{x}_t$ is a minimizer by convexity; hence, without loss of generality, we assume that $\nabla f(\boldsymbol{x}_t) \neq 0$ for all t.

Throughout this paper, we assume that the objective f is convex and Hölder smooth, which are standard assumptions in the optimization literature (Nesterov, 2015).

Assumption 1 (Convexity)**.** *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, it holds that*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

Assumption 2 ((ν, M_ν)-Hölder Smoothness)**.** *There exist $\nu \in [0, 1]$ and $M_\nu > 0$ such that, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq M_\nu \|\boldsymbol{x} - \boldsymbol{y}\|^\nu. \tag{3}$$

Hölder smoothness interpolates between non-smooth and smooth objectives. For $\nu = 1$, Eq. (3) is the standard M_1-smoothness assumption, i.e., Lipschitz continuity of the gradient, which allows the gradient norm to grow without bound, as for quadratic functions. For $\nu = 0$, Eq. (3) is essentially the bounded-gradient assumption: it holds with $M_0 = 2G$ whenever $\|\nabla f(\boldsymbol{x})\| \leq G$ for all $\boldsymbol{x}$, and, conversely, taking $\boldsymbol{y} = \boldsymbol{x}^\star$ gives $\|\nabla f(\boldsymbol{x})\| \leq M_0$. In this case, f need not be differentiable, and $\nabla f(\boldsymbol{x})$ denotes an arbitrary subgradient of f at $\boldsymbol{x}$, with $\nabla f(\boldsymbol{x}^\star) = 0$. For $\nu = 0$, our proofs use only the subgradient inequality of Assumption 1, the bound $\|\nabla f(\boldsymbol{x})\| \leq M_0$, and the upper bound $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + M_0 \|\boldsymbol{y} - \boldsymbol{x}\|$, all of which hold for subgradients of convex functions with $M_0/2$-bounded subgradients; hence our results for $\nu = 0$ apply to the normalized subgradient method on non-smooth convex functions, which is the setting of Zamani & Glineur (2025).

3.1 Existing Analysis of Normalized Gradient Descent

As discussed in Sec. 2, most existing analyses of normalized gradient descent bound either the best iterate or a weighted average of the iterates. Under Hölder smoothness, the following two guarantees are known (see also Table 1).

Proposition 1 (Grimmer (2019))**.** *Suppose that Assumptions 1 and 2 hold. Then, for any $T \geq 1$, normalized gradient descent with $\tilde{\eta}_t = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ satisfies:*

$$\min_{t \in \{1,2,\ldots,T\}} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \mathcal{O}\left(M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} \right). \tag{4}$$

Proposition 2 (Orabona (2023))**.** *Suppose that Assumptions 1 and 2 hold. Then, for any $T \geq 1$, normalized gradient descent with $\tilde{\eta}_t = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ satisfies:*

$$f(\tilde{\boldsymbol{x}}_T) - f(\boldsymbol{x}^\star) \leq \mathcal{O}\left(M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} \right), \tag{5}$$

where $\tilde{\boldsymbol{x}}_T = \frac{1}{\sum_{t=1}^T \frac{1}{\|\nabla f(\boldsymbol{x}_t)\|}} \sum_{t=1}^T \frac{\boldsymbol{x}_t}{\|\nabla f(\boldsymbol{x}_t)\|}$.

Proposition 1 bounds the best iterate, i.e., the iterate with the smallest function value, which can be identified only by evaluating f at every iterate. Proposition 2 bounds a weighted average of the iterates, with weights proportional to the effective stepsizes η_t. In practice, however, normalized gradient descent typically returns the last iterate $\boldsymbol{x}_T$, which neither result covers. To the best of our knowledge, the only existing last-iterate guarantee for normalized gradient descent is that of Zamani & Glineur (2025), which is restricted to $\nu = 0$. In the next section, we establish last-iterate guarantees for the whole Hölder family $\nu \in [0, 1]$.

4 Normalized Gradient Descent with a Constant Stepsize

In this section, we consider normalized gradient descent with a constant stepsize, i.e., $\tilde{\eta}_t = \tilde{\eta}$ for all t. All proofs are deferred to App. C.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Then, for any $T \geq 1$, normalized gradient descent with $\tilde{\eta}_t = \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ satisfies*

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \mathcal{O}\left(M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} (1 + \log(T))^{1+\nu} \right).$$

Discussion. Theorem 1 provides a last-iterate guarantee for normalized gradient descent under Hölder smoothness. As discussed in Sec. 2, Zamani & Glineur (2025) analyzed the last iterate only under the bounded-gradient assumption, i.e., for $\nu = 0$. Theorem 1 recovers their result up to absolute constants as a special case and extends it to the whole Hölder family $\nu \in [0, 1]$, which, in particular, includes smooth objectives with unbounded gradients. See the detailed comparison in Table 1.

Compared with the best-iterate and weighted-average guarantees in Propositions 1 and 2, the bound in Theorem 1 is worse by a factor of $(1 + \log(T))^{1+\nu}$. For $\nu = 0$, this overhead is not an artifact of

the analysis: Zamani & Glineur (2025) established a lower bound for normalized gradient descent with the constant stepsize $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$ that matches Theorem 1 in this case, so the logarithmic factor cannot be removed in general for this stepsize. For $\nu > 0$, no such lower bound is currently known. In Sec. 4.2, we therefore study the smooth case $\nu = 1$ numerically and find that the last iterate of normalized gradient descent with a constant stepsize converges slower than $\mathcal{O}(1/T)$ in the worst case, over the range of horizons we can solve.

This behavior is particularly notable in the smooth case. For convex and smooth objectives, gradient descent with a constant stepsize attains the rate $\mathcal{O}(1/T)$ for the last iterate (Nesterov, 2018), which is the same order as the best-iterate rate of normalized gradient descent in Proposition 1. In contrast, the numerical results in Sec. 4.2 indicate that normalized gradient descent does not attain this rate for the last iterate. Thus, normalization can degrade the last-iterate behavior relative to gradient descent, even though it does not affect the best-iterate rate.

4.1 Proof Sketch

We now outline the proof of Theorem 1, explaining where the logarithmic factor comes from and why the case $\nu > 0$ requires an argument different from that of Zamani & Glineur (2025).

Key lemma. Both of our main results rely on the following lemma, which is a variant of Lemma 2.1 of Zamani & Glineur (2025) and of an argument in the proof of Lemma 4.1 of Liu & Zhou (2024b). Its proof compares each iterate $\boldsymbol{x}_t$ with a reference point $\boldsymbol{z}_t$ that is a convex combination of $\boldsymbol{x}^\star$ and the previous iterates, with weights determined by the sequence $\{v_t\}_t$.

Lemma 1. *Suppose that Assumption 1 holds. Then, for any sequence $\{v_t\}_{t=0}^T$ such that $0 < v_0 \leq v_1 \leq \cdots \leq v_T$, normalized gradient descent satisfies*

$$2\sum_{t=1}^{T}\left[\eta_t v_t - \left(\sum_{s=t}^{T}\eta_s\right)(v_t - v_{t-1})\right](f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) \leq v_0\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^{T} v_t\tilde{\eta}_t^2.$$

Reduction to two terms. To simplify the notation, we write $\eta_t \coloneqq \frac{\tilde{\eta}_t}{\|\nabla f(\boldsymbol{x}_t)\|}$. To isolate the last iterate, we choose

$$v_T \coloneqq \frac{1}{\eta_T}, \quad v_t \coloneqq \frac{1}{\sum_{s=t+1}^{T}\eta_s} \quad \text{for all } t \in \{0, 1, \ldots, T-1\}. \tag{6}$$

This sequence is nondecreasing, and with this choice the coefficient of $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)$ on the left-hand side of Lemma 1 vanishes for every $t < T$ and equals 2 for $t = T$. Hence, for $\tilde{\eta}_t \coloneqq \tilde{\eta}$, we obtain the following by Lemma 1:

$$2\left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right) \leq v_0\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \tilde{\eta}^2\sum_{t=1}^{T} v_t. \tag{7}$$

Hence, it remains to bound v_0 and $\tilde{\eta}^2\sum_{t=1}^{T} v_t$ from above, and both quantities are controlled by the gradient norms at the iterates. For v_0, dropping the term $s = T$ and applying Lemma 4, a consequence of Assumption 2 that bounds $\|\nabla f(\boldsymbol{x})\|$ by $2M_\nu^{\frac{1}{1+\nu}}(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}}$, we get

$$v_0 = \frac{1}{\tilde{\eta}\sum_{s=1}^{T}\frac{1}{\|\nabla f(\boldsymbol{x}_s)\|}} \leq \frac{1}{\tilde{\eta}\sum_{s=1}^{T-1}\frac{1}{\|\nabla f(\boldsymbol{x}_s)\|}} \leq \frac{2M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}}\left(\sum_{s=1}^{T-1}\left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{-\frac{\nu}{1+\nu}}\right)^{-1}. \tag{8}$$

Similarly, using the inequality between the harmonic and the arithmetic mean, Lemma 4, the definition of v_T, and a change of the order of summation, we get

$$\begin{aligned}\tilde{\eta}^2 \sum_{t=1}^{T} v_t &= \tilde{\eta}^2 v_T + \tilde{\eta} \sum_{t=1}^{T-1} \frac{1}{\sum_{s=t+1}^{T} \frac{1}{\|\nabla f(\boldsymbol{x}_s)\|}} \\ &\leq \tilde{\eta}^2 v_T + \tilde{\eta} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T} \|\nabla f(\boldsymbol{x}_s)\| \qquad (9) \\ &\leq \tilde{\eta}^2 v_T + 2M_\nu^{\frac{1}{1+\nu}} \tilde{\eta} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T} (f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}} \\ &\leq 6M_\nu^{\frac{1}{1+\nu}} \tilde{\eta} \, (f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}} + 2M_\nu^{\frac{1}{1+\nu}} \tilde{\eta} \sum_{s=2}^{T-1} (f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2}.\end{aligned}$$

The details are given in App. C. When $\nu = 0$, the factors $(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}}$ are equal to one, and the last-iterate rate follows directly from Eqs. (7) to (9), as in Zamani & Glineur (2025). For $\nu \in (0, 1]$, however, the bounds on v_0 and $\tilde{\eta}^2 \sum_{t=1}^{T} v_t$ depend on the suboptimality gaps $f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)$ of the earlier iterates, i.e., on the very quantities we want to bound. This is what makes the case $\nu > 0$ harder: we resolve it by proving the following bound for *every* iterate by induction on t.

Lemma 2. *Suppose that Assumptions 1 and 2 hold, and let $\tilde{\eta}_t = \tilde{\eta} > 0$ for all t. Then, for every $t \geq 1$ such that $\tilde{\eta} \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{t}}$, normalized gradient descent satisfies*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq M_\nu \left(1200 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} t} \right)^{1+\nu} + 200 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{t} \right)^{\frac{1+\nu}{2}} \right) (1 + \log(t))^{1+\nu}. \quad (10)$$

Since $\tilde{\eta} = \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$ satisfies $\tilde{\eta} \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{t}$ for every $t \leq T$, Theorem 1 follows from Lemma 2 with $t = T$, with the absolute constant 1400.

Proof Sketch of Lemma 2. We sketch the induction step, which also shows where the logarithmic factor comes from; the base case $t = 1$ follows from Lemma 3.

Assume that Eq. (10) holds for all $1 \leq t \leq T-1$; we show that it holds for $t = T$ whenever $\tilde{\eta} \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$. By Eq. (7), it suffices to bound v_0 and $\tilde{\eta}^2 \sum_{t=1}^{T} v_t$. Combining Eq. (8) with the induction hypothesis (which applies to every $s \leq T-1$ because $\tilde{\eta} \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{s}$) and elementary manipulations, we obtain

$$v_0 \leq 320(1 + \log(T))^\nu M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{2\nu}}{\tilde{\eta}^{1+\nu} T^{1+\nu}} \right).$$

The factor $(1 + \log(T))^\nu$ is inherited from the induction hypothesis through $(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}}$; the bound on v_0 itself does not introduce any additional logarithmic factor.

The additional logarithmic factor arises in the bound on $\tilde{\eta}^2 \sum_{t=1}^{T} v_t$, so we discuss it in more detail. By Eq. (9), it suffices to bound

$$\begin{aligned}&\sum_{s=2}^{T-1} (f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2} \\ &= \underbrace{\sum_{s=2}^{\lfloor \frac{T-1}{2} \rfloor} (f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2}}_{\mathcal{T}_1} + \underbrace{\sum_{s=\lfloor \frac{T-1}{2} \rfloor + 1}^{T-1} (f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star))^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2}}_{\mathcal{T}_2}.\end{aligned}$$

In $\mathcal{T}_1$, the index s is at most $\frac{T-1}{2}$, so $T - t \geq \frac{T}{2}$ for all $t \leq s-1$ and the inner sum is at most $\frac{4s}{T^2}$; combining this with Eq. (10) for $s \leq T-1$, we obtain, without any additional logarithmic factor,

$$\mathcal{T}_1 \leq 200 M_\nu^{\frac{\nu}{1+\nu}} (1 + \log(T))^\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{2\nu}}{\tilde{\eta}^\nu T^\nu} \right).$$

In $\mathcal{T}_2$, the index s can be close to T, and then the inner sum $\sum_{t=1}^{s-1} \frac{1}{(T-t)^2} \leq \frac{1}{T-s}$ is no longer small; summing $\frac{1}{T-s}$ over s produces a harmonic sum, i.e., an additional logarithmic factor. Specifically, using $s \geq \lfloor \frac{T-1}{2} \rfloor + 1$, $T \geq 2$, and Eq. (10), we have

$$\begin{aligned}
\mathcal{T}_2 &\leq M_\nu^{\frac{\nu}{1+\nu}} (1+\log(T))^\nu \sum_{s=\lfloor \frac{T-1}{2} \rfloor}^{T-1} \left(70 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T} \right)^\nu + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{\nu}{2}} \right) \sum_{t=1}^{s-1} \frac{1}{(T-t)^2} \\
&\leq M_\nu^{\frac{\nu}{1+\nu}} (1+\log(T))^\nu \sum_{s=\lfloor \frac{T-1}{2} \rfloor}^{T-1} \left(70 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T} \right)^\nu + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{\nu}{2}} \right) \frac{1}{T-s} \qquad (11) \\
&\leq M_\nu^{\frac{\nu}{1+\nu}} (1+\log(T))^{1+\nu} \left(70 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T} \right)^\nu + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{\nu}{2}} \right).
\end{aligned}$$

The last step uses $\sum_{s=\lfloor \frac{T-1}{2} \rfloor}^{T-1} \frac{1}{T-s} \leq 1 + \log(T)$. This harmonic sum is the source of the additional logarithmic factor, and it is precisely the reason why the induction hypothesis Eq. (10) must itself contain a logarithmic factor. Combining the bounds on v_0, $\mathcal{T}_1$, and $\mathcal{T}_2$ with Eqs. (7) and (9) and using $\tilde{\eta} \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$, we obtain

$$\begin{aligned}
f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq{}& \frac{4 M_\nu^{\frac{1}{1+\nu}} \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T\tilde{\eta}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \right)^{\frac{\nu}{1+\nu}} \\
&+ M_\nu (1+\log(T))^{1+\nu} \left(430 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T} \right)^{1+\nu} + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} \right),
\end{aligned}$$

and Lemma 5, a consequence of Young's inequality, yields Eq. (10) for $t = T$.

4.2 Numerical Evaluation

In this section, we use the performance estimation problem (PEP) (Drori & Teboulle, 2014; Taylor et al., 2017) to examine numerically whether the logarithmic overhead in Theorem 1 is an artifact of our analysis. For $\nu = 0$, it is not: the lower bound of Zamani & Glineur (2025) matches Theorem 1 for the stepsize $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$. For $\nu > 0$, no lower bound is known, so we focus on the smooth case $\nu = 1$ and ask whether the last iterate attains the $\mathcal{O}(1/T)$ rate of the best iterate.

PEP computes the exact worst-case performance of a first-order method over a class of functions by solving a semidefinite program. Since we are interested in the dependence on T, we set $M_\nu = 1$ and $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\| = 1$. The worst-case last-iterate suboptimality of normalized gradient descent after T iterations is then the optimal value of

$$\begin{aligned}
\max_f \; & f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \quad \text{s.t.} \quad \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\| = 1, \\
& \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \qquad (12) \\
& f \text{ satisfies Assumptions 1 and 2 with } M_\nu = 1 \text{ and } \nu = 1.
\end{aligned}$$

Because the normalization $\nabla f(\boldsymbol{x}_t)/\|\nabla f(\boldsymbol{x}_t)\|$ is nonlinear in $\nabla f(\boldsymbol{x}_t)$, Eq. (12) cannot be cast as a semidefinite program directly. We therefore fix the gradient norms $\{g_t\}_t$ in advance and solve

$$\begin{aligned}
\max_f \; & f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \quad \text{s.t.} \quad \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\| = 1, \\
& \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} \frac{\nabla f(\boldsymbol{x}_t)}{g_t}, \qquad (13) \\
& \|\nabla f(\boldsymbol{x}_t)\| = g_t \text{ for all } 1 \leq t \leq T-1, \\
& f \text{ satisfies Assumptions 1 and 2 with } M_\nu = 1 \text{ and } \nu = 1,
\end{aligned}$$

which can be formulated as a semidefinite program: the update is now linear in $\nabla f(\boldsymbol{x}_t)$, and the norm constraints are linear in the Gram matrix of the gradients; see Drori & Teboulle (2014); Taylor

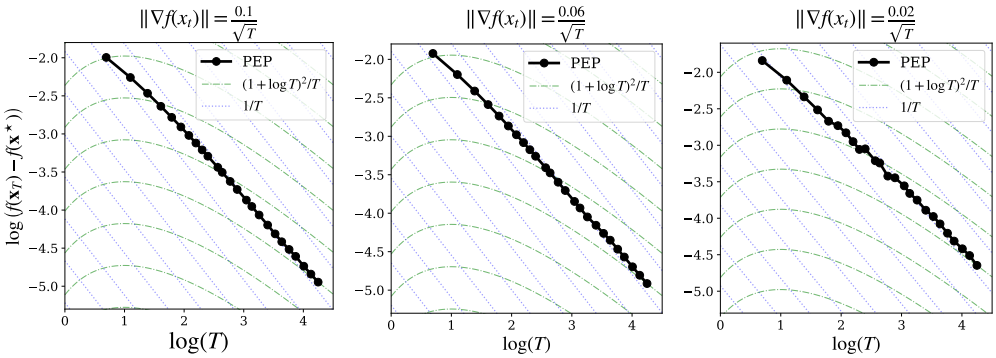

Figure 1: Optimal value of Eq. (13), i.e., the worst-case $f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)$ over convex 1-smooth functions with $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\| = 1$ and prescribed gradient norms $g_t = c/\sqrt{T}$, as a function of T (log–log scale) for $c = 0.1$ (left), $c = 0.06$ (middle), and $c = 0.02$ (right). The dotted and dash-dotted lines show the reference decays $1/T$ and $(1 + \log T)^2/T$ up to constant factors. In all three cases, the observed decay is slower than $1/T$.

et al. (2017) for the details. We solve it with PEPit (Goujaud et al., 2024). Every feasible point of Eq. (13) is feasible for Eq. (12); hence, for any choice of $\{g_t\}_t$, the optimal value of Eq. (13) is a lower bound on the worst-case suboptimality of normalized gradient descent after T iterations, and maximizing it over $\{g_t\}_t$ would recover Eq. (12). Since this maximization is intractable, and since a single sequence suffices to exhibit slow convergence, we tried several choices of $\{g_t\}_t$ and report the results for $g_t = c/\sqrt{T}$ with $c \in \{0.1, 0.06, 0.02\}$, for which the observed decay is slower than $1/T$. Other choices of $\{g_t\}_t$ may lead to even larger worst-case values.

Figure 1 shows the results. For all three values of c, the worst-case suboptimality decays slower than $\mathcal{O}(1/T)$ over the range of horizons that we could solve, but faster than the $\mathcal{O}(\log^2(T)/T)$ bound of Theorem 1. The latter does not imply that Theorem 1 is loose, because $g_t \propto 1/\sqrt{T}$ need not be the worst choice of the gradient norms. We stress that these experiments cover a limited range of T and a hand-picked family $\{g_t\}_t$: they indicate that the last iterate of normalized gradient descent with a constant stepsize does not attain the $\mathcal{O}(1/T)$ rate of the best iterate, but they neither identify the exact worst-case rate nor prove that the overhead is logarithmic. Determining the tightness of Theorem 1 for $\nu > 0$ is left for future work.

5 Normalized Gradient Descent with a Linearly Decreasing Stepsize

Section 4 shows that, with a constant stepsize, the last iterate of normalized gradient descent suffers a logarithmic overhead which, at least for $\nu = 0$, is unavoidable. In this section, we show that this overhead disappears if the stepsize linearly decreases. Specifically, we consider the linearly decreasing stepsize

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \underbrace{\eta\left(1 - \frac{t}{T+1}\right)}_{\tilde{\eta}_t} \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \tag{14}$$

where $\eta > 0$ is the *baseline* stepsize, i.e., the maximum stepsize over the run. This is the linear-decay schedule widely used for training deep neural networks (Devlin et al., 2019; Liu et al., 2021; Defazio et al., 2024a); note that it depends only on the iteration counter t and on the horizon T, and not on the iterates or gradients.

The following theorem gives the last-iterate rate of normalized gradient descent with the stepsize Eq. (14) and $\eta = \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|/\sqrt{T}$. The proof is deferred to App. D.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then, for any $T \geq 1$, normalized gradient descent with $\tilde{\eta}_t = (1 - \frac{t}{T+1})\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ satisfies*

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \mathcal{O}\left(M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}\right).$$

Discussion. Theorem 2 shows that the linearly decreasing stepsize removes the logarithmic overhead of Theorem 1: the last iterate attains the same order $\mathcal{O}\big(M_\nu(\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2/T)^{(1+\nu)/2}\big)$ as the best-iterate and weighted-average guarantees in Propositions 1 and 2, and the stepsize requires no knowledge of ν or M_ν. For $\nu = 0$, this recovers, up to absolute constants, the result of Zamani & Glineur (2025), who showed that the normalized subgradient method with the linearly decreasing stepsize $\tilde{\eta}_t = \|\boldsymbol{x}_1-\boldsymbol{x}^\star\|(1-\frac{t}{T})/\sqrt{T}$ attains the last-iterate rate $\mathcal{O}(1/\sqrt{T})$ for convex functions with bounded subgradients; their schedule differs from Eq. (14) only by the shift $T \to T+1$. Theorem 2 extends this guarantee to the whole Hölder family $\nu \in [0, 1]$.

5.1 Proof Sketch

We now explain why a linearly decreasing stepsize removes the logarithmic factor. As discussed in Sec. 4.1, this factor comes from the harmonic sum $\sum_s \frac{1}{T-s}$ in the bound on $\tilde{\eta}^2 \sum_t v_t$ (see Eq. (11)), which is dominated by the iterates s close to T. A stepsize that decreases towards the end of the run reduces the weight of exactly these iterates.

Similar to Sec. 4.1, we choose $\{v_t\}_t$ as in Eq. (6), so that Lemma 1 gives

$$2\left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right) \le v_0\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^{T} v_t\tilde{\eta}_t^2,$$

and it remains to bound v_0 and $\sum_{t=1}^T v_t\tilde{\eta}_t^2$ from above. For the linear schedule, we have the identity $\sum_{s=t+1}^T \tilde{\eta}_s = \tilde{\eta}_t(T-t)/2$, and the Cauchy–Schwarz inequality then yields $v_t\tilde{\eta}_t^2 \le \frac{4}{(T-t)^2}\sum_{s=t+1}^T \tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|$ for every $t < T$; the sums $\sum_{s=t+1}^T \tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|$ are in turn controlled by the suboptimality gaps $f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)$ (Lemma 9).

For the constant stepsize, we bounded $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)$ for every t, and the bound necessarily contained a logarithmic factor. For the linearly decreasing stepsize, we instead show that $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)$ is bounded by $\mathcal{O}(T^{-\frac{1+\nu}{2}})$ on a sequence of iterates that becomes denser towards the end of the run. Specifically, we partition the iterations into the blocks

$$\mathcal{I}_j \coloneqq \left\{ t \in \{1,\dots,T\} \;\middle|\; 2^{-(j+1)} < 1 - \frac{t}{T+1} \le 2^{-j} \right\} \quad \text{for all } j \in \{0,\dots,J\},$$

where $J \coloneqq \lceil \log_2(\frac{T+1}{2\sqrt{T}}) \rceil$, so that the stepsize is halved from one block to the next and $|\mathcal{I}_j| \approx (T+1)/2^{j+1}$. This decomposition is similar in spirit to that of Jain et al. (2019), who halve the stepsize over phases of geometrically decreasing length to remove the logarithmic factor from the last-iterate rate of SGD. In contrast, our stepsize decreases linearly, and the blocks $\mathcal{I}_j$ are used only in the analysis. We show that each block contains an iterate $s_j \in \mathcal{I}_j$ with

$$f(\boldsymbol{x}_{s_j}) - f(\boldsymbol{x}^\star) \le 23104 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}.$$

See Lemma 8 in details. In words, the best-iterate guarantee of Proposition 1 ensures that the rate $\mathcal{O}(T^{-\frac{1+\nu}{2}})$ is attained at least once during the run; with the linearly decreasing stepsize, it is attained once in every block, and the blocks shrink geometrically towards the end of the run. Since the iterates close to T are the ones that generate the harmonic sum, having such good iterates in every block allows us to bound v_0 and $\sum_t v_t\tilde{\eta}_t^2$ without a logarithmic factor, which yields Theorem 2. The last at most $\sqrt{T}$ iterations, which are not covered by the blocks, have stepsize $\tilde{\eta}_t$ at most $\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|/(T+1)$, and their total contribution is negligible.

6 Conclusion

We studied the last-iterate convergence of normalized gradient descent for convex and Hölder-smooth objectives. With a constant stepsize, the last iterate achieves $\mathcal{O}\big((\log^2(T)/T)^{(1+\nu)/2}\big)$, which is slower than the known best-iterate and weighted-average rates by a logarithmic factor; this factor is unavoidable for $\nu = 0$, and our PEP-based numerical results indicate that the last iterate is also slower than $\mathcal{O}(1/T)$ in the smooth case. With a linearly decreasing stepsize, the logarithmic factor disappears and the last iterate attains the rate $\mathcal{O}\big(T^{-(1+\nu)/2}\big)$ without knowledge of the Hölder parameters.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and proofreading of the manuscript, including improving grammar, clarity, and readability. We have not used generative AI tools to develop the main research ideas, derive theoretical results, or write the code in our experiments. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *arXiv*, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *arXiv*, 2020.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. In *Mathematical Programming*, 2021.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, 2020.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. In *arXiv*, 2024.

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. In *arXiv*, 2024a.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. In *Advances in Neural Information Processing Systems*, 2024b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. In *Mathematical Programming*, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Journal of Machine Learning Research*, 2011.

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. High probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. In *arXiv*, 2024.

Baptiste Goujaud, Céline Moucer, François Glineur, Julien Hendrickx, Adrien Taylor, and Aymeric Dieuleveut. PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python. In *Mathematical Programming Computation*, 2024.

Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. In *SIAM Journal on Optimization*, 2019.

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

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Conference on Learning Theory*, 2019.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Diederik P. Kingma and Jim Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Guy Kornowski and Ohad Shamir. Gradient descent's last iterate is often (slightly) suboptimal. In *arXiv*, 2026.

Xiaoyu Li, Mingrui Liu, and Francesco Orabona. On the last iterate convergence of momentum methods. In *International Conference on Algorithmic Learning Theory*, 2022.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. In *arXiv*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In *International Conference on Machine Learning*, 2024a.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. In *Advances in Neural Information Processing Systems*, 2024.

Yu Nesterov. Universal gradient methods for convex optimization problems. In *Mathematical Programming*, 2015.

Yurii Nesterov. Lectures on convex optimization. In *Springer*, 2018.

Francesco Orabona. Normalized gradients for all. In *arXiv*, 2023.

Margarita Preobrazhenskaia, Makar Sidorov, Igor Preobrazhenskii, and Eduard Gorbunov. Last iterate convergence of adagrad-norm for convex non-smooth optimization. In *arXiv*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision*, 2015.

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. In *International Conference on Machine Learning*, 2025.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, 2013.

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

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, 2025.

Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 2025.

Moslem Zamani and François Glineur. Exact convergence rate of the subgradient method by using polyak step size. In *arXiv*, 2024.

A Importance of Convergence Analysis in the Convex Setting

In this paper, we assume that the objective is convex. In practical applications, most loss functions are non-convex; nevertheless, analysis in the convex setting remains highly important. One reason is that convergence analyses under non-convex assumptions are often overly pessimistic. Since optimization theory typically studies worst-case functions, if one allows arbitrary non-convex functions, the resulting guarantees can become too conservative to accurately reflect the behavior observed in practical problems. For example, under the assumptions of non-convexity and smoothness, gradient descent is known to be an optimal first-order optimization method for finding approximate stationary points, and its convergence rate cannot be further improved in general (Carmon et al., 2021). Furthermore, several recent works report that predictions of convex optimization theory match the behavior of neural network training surprisingly well, e.g., regarding learning-rate schedules (Defazio et al., 2024b; Schaipp et al., 2025). We believe that our analysis of normalized gradient descent in the convex setting provides useful insight beyond the convex regime and may serve as a basis for understanding its empirical effectiveness in broader non-convex problems.

B Useful Lemmas

Lemma 3 (Nesterov (2015))**.** *Suppose that Assumption 2 holds. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, it holds that*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{M_\nu}{1+\nu} \|\boldsymbol{x} - \boldsymbol{y}\|^{1+\nu}.$$

Lemma 4 (Lemma A.4 in Gorbunov et al. (2024))**.** *Suppose that Assumption 2 holds. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$, it holds that*

$$\|\nabla f(\boldsymbol{x})\| \leq \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}},$$

where for $\nu = 0$, we use $\lim_{\nu \to 0} (\frac{1+\nu}{\nu})^{\frac{\nu}{1+\nu}} = 1$.

Lemma 5. *For any $x, a, b \geq 0$ and $\nu \in [0, 1]$, if*

$$x \leq a x^{\frac{\nu}{1+\nu}} + b,$$

then it holds that

$$x \leq a^{1+\nu} + 2b.$$

Proof. From Young's inequality, we have

$$a x^{\frac{\nu}{1+\nu}} \leq \frac{\nu}{1+\nu} x + \frac{a^{1+\nu}}{1+\nu}.$$

Substituting this bound into the assumption gives $\frac{x}{1+\nu} \leq \frac{a^{1+\nu}}{1+\nu} + b$, i.e., $x \leq a^{1+\nu} + (1+\nu)b \leq a^{1+\nu} + 2b$. □

Lemma 1. *Suppose that Assumption 1 holds. Then, for any sequence $\{v_t\}_{t=0}^T$ such that $0 < v_0 \leq v_1 \leq \cdots \leq v_T$, normalized gradient descent satisfies*

$$2 \sum_{t=1}^T \left[\eta_t v_t - \left(\sum_{s=t}^T \eta_s \right) (v_t - v_{t-1}) \right] \left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\right) \leq v_0 \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^T v_t \tilde{\eta}_t^2.$$

Proof. Let $\boldsymbol{z}_0 = \boldsymbol{x}^\star$ and $\boldsymbol{z}_t$ be defined recursively as follows:

$$\boldsymbol{z}_t = \left(1 - \frac{v_{t-1}}{v_t}\right) \boldsymbol{x}_t + \frac{v_{t-1}}{v_t} \boldsymbol{z}_{t-1}.$$

We have

$$\begin{aligned}\|\boldsymbol{x}_{t+1}-\boldsymbol{z}_t\|^2 &= \|\boldsymbol{x}_t-\boldsymbol{z}_t\|^2 - 2\eta_t\langle\nabla f(\boldsymbol{x}_t),\boldsymbol{x}_t-\boldsymbol{z}_t\rangle + \tilde{\eta}_t^2\\ &\le \|\boldsymbol{x}_t-\boldsymbol{z}_t\|^2 - 2\eta_t\left(f(\boldsymbol{x}_t)-f(\boldsymbol{z}_t)\right) + \tilde{\eta}_t^2\\ &\le \left(\frac{v_{t-1}}{v_t}\right)\|\boldsymbol{x}_t-\boldsymbol{z}_{t-1}\|^2 - 2\eta_t\left(f(\boldsymbol{x}_t)-f(\boldsymbol{z}_t)\right) + \tilde{\eta}_t^2,\end{aligned}$$

where we use Assumption 1 for the first inequality and $v_{t-1}\le v_t$ for the last inequality. Thus, we get

$$2\eta_t v_t\left(f(\boldsymbol{x}_t)-f(\boldsymbol{z}_t)\right) \le v_{t-1}\|\boldsymbol{x}_t-\boldsymbol{z}_{t-1}\|^2 - v_t\|\boldsymbol{x}_{t+1}-\boldsymbol{z}_t\|^2 + v_t\tilde{\eta}_t^2.$$

Unrolling the definition of $\boldsymbol{z}_t$ gives $\boldsymbol{z}_t = \sum_{s=1}^t \frac{v_s-v_{s-1}}{v_t}\boldsymbol{x}_s + \frac{v_0}{v_t}\boldsymbol{x}^\star$, a convex combination since $\{v_t\}_t$ is nondecreasing and positive; hence, by the convexity of f, we have

$$f(\boldsymbol{z}_t) \le \sum_{s=1}^t \frac{v_s-v_{s-1}}{v_t} f(\boldsymbol{x}_s) + \frac{v_0}{v_t} f(\boldsymbol{x}^\star).$$

Multiplying this inequality by $2\eta_t v_t$, summing the resulting bounds and the previous display over $t=1,\dots,T$, exchanging the order of summation, and using $\boldsymbol{z}_0=\boldsymbol{x}^\star$ and $\|\boldsymbol{x}_{T+1}-\boldsymbol{z}_T\|^2\ge 0$, we obtain

$$\begin{aligned}2\sum_{t=1}^T\left[\eta_t v_t - \left(\sum_{s=t}^T\eta_s\right)(v_t-v_{t-1})\right]\left(f(\boldsymbol{x}_t)-f(\boldsymbol{x}^\star)\right) &\le 2\sum_{t=1}^T \eta_t v_t\left(f(\boldsymbol{x}_t)-f(\boldsymbol{z}_t)\right)\\ &\le v_0\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2 + \sum_{t=1}^T v_t\tilde{\eta}_t^2.\end{aligned}$$

□

C Proof of Theorem 1

Lemma 2. *Suppose that Assumptions 1 and 2 hold, and let $\tilde{\eta}_t=\tilde{\eta}>0$ for all t. Then, for every $t\ge 1$ such that $\tilde{\eta}\le\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{t}}$, normalized gradient descent satisfies*

$$f(\boldsymbol{x}_t)-f(\boldsymbol{x}^\star) \le M_\nu\left(1200\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{\tilde{\eta}t}\right)^{1+\nu} + 200\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{t}\right)^{\frac{1+\nu}{2}}\right)(1+\log(t))^{1+\nu}. \quad (10)$$

Proof. We prove the statement by induction on t. For $t=1$, Lemma 3 with $\boldsymbol{x}=\boldsymbol{x}^\star$ and $\nabla f(\boldsymbol{x}^\star)=0$ gives

$$f(\boldsymbol{x}_1)-f(\boldsymbol{x}^\star) \le \frac{M_\nu}{1+\nu}\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{1+\nu},$$

which is at most the right-hand side of Eq. (10) for $t=1$ since $200\ge\frac{1}{1+\nu}$. Thus, the statement holds for $t=1$.

Now let $T\ge 2$, assume that $\tilde{\eta}\le\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$, and assume that Eq. (10) holds for every $t\le T-1$; note that the induction hypothesis is applicable to every $t\le T-1$, because $\tilde{\eta}\le\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\le\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{t}}$. We show that Eq. (10) holds for $t=T$. Let $\{v_t\}_{t=0}^T$ be defined as follows:

$$\begin{aligned}v_T &:= \frac{1}{\eta_T}\\ v_t &:= \frac{1}{\sum_{s=t+1}^T\eta_s} \quad (\text{for all } t=0,1,\dots,T-1).\end{aligned}$$

Note that $\{v_t\}_{t=0}^T$ is positive and nondecreasing (in fact, $v_{T-1}=v_T$). In this case, we have

$$\eta_T v_T = 1,$$

and for any $t \in \{1, 2, \dots, T-1\}$, we have

$$\eta_t v_t - \left(\sum_{s=t}^{T} \eta_s\right)(v_t - v_{t-1}) = -\left(\sum_{s=t+1}^{T} \eta_s\right) v_t + \left(\sum_{s=t}^{T} \eta_s\right) v_{t-1} = 0.$$

Thus, from Lemma 1, we have

$$2\left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right) \le v_0 \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \tilde{\eta}^2 \sum_{t=1}^{T} v_t.$$

We first bound v_0. Dropping the term $s = T$, we have

$$v_0 = \frac{1}{\tilde{\eta} \sum_{s=1}^{T} \frac{1}{\|\nabla f(\boldsymbol{x}_s)\|}} \le \frac{1}{\tilde{\eta} \sum_{s=1}^{T-1} \frac{1}{\|\nabla f(\boldsymbol{x}_s)\|}}.$$

For any $s \le T-1$, the induction hypothesis Eq. (10) and the inequality $\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{s}\right)^{\frac{1+\nu}{2}} \le \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} s}\right)^{1+\nu}$, which follows from $\tilde{\eta} \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{s}}$, give $f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star) \le 1400 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} s}\right)^{1+\nu} (1 + \log(s))^{1+\nu}$. Combining this with Lemma 4, $\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \le 2$, and $1400^{\frac{\nu}{1+\nu}} \le \sqrt{1400} < 40$, it holds for any $s \le T-1$ that

$$\|\nabla f(\boldsymbol{x}_s)\| \le 80 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} s}\right)^{\nu} (1 + \log(s))^{\nu}.$$

Applying the above inequality and $\sum_{s=1}^{T-1} s^\nu \ge \frac{T^{1+\nu}}{4}$ for $T \ge 2$, we obtain

$$\begin{aligned} v_0 &\le \frac{80 M_\nu \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{2\nu}}{\tilde{\eta}^{1+\nu}} (1 + \log(T))^{\nu} \frac{1}{\sum_{s=1}^{T-1} s^\nu} \\ &\le 320 (1 + \log(T))^{\nu} M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{2\nu}}{\tilde{\eta}^{1+\nu} T^{1+\nu}}\right). \end{aligned}$$

Next, we bound $\sum_{t=1}^{T} v_t$:

$$\begin{aligned} \sum_{t=1}^{T} v_t &= v_T + \frac{1}{\tilde{\eta}} \sum_{t=1}^{T-1} \frac{1}{\sum_{s=t+1}^{T} \frac{1}{\|\nabla f(\boldsymbol{x}_s)\|}} \\ &\le v_T + \frac{1}{\tilde{\eta}} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T} \|\nabla f(\boldsymbol{x}_s)\| \\ &\le v_T + \frac{1}{\tilde{\eta}} \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \\ &\le \frac{M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \\ &\quad + \frac{M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \\ &\le \frac{M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \left(1 + \sum_{t=1}^{T-1} \frac{1}{(T-t)^2}\right) \\ &\quad + \frac{M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T-1} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \\ &\le \frac{6 M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} + \frac{2 M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \sum_{s=t+1}^{T-1} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \\ &= \frac{6 M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} + \frac{2 M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \sum_{s=2}^{T-1} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2}, \end{aligned}$$

where we use the fact that the harmonic mean is smaller than or equal to the arithmetic mean for the first inequality, Lemma 4 (a consequence of Assumption 2) for the second and third inequalities, and $\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \leq 2$ together with $1 + \sum_{t=1}^{T-1} \frac{1}{(T-t)^2} \leq 1 + \frac{\pi^2}{6} \leq 3$ for the fifth inequality; the last step exchanges the order of summation. Then, we have

$$\begin{aligned}
&\frac{2M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \sum_{s=2}^{T-1} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2} \\
&= \frac{2M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \underbrace{\sum_{s=2}^{\lfloor \frac{T-1}{2} \rfloor} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2}}_{\mathcal{T}_1} \\
&\quad + \frac{2M_\nu^{\frac{1}{1+\nu}}}{\tilde{\eta}} \underbrace{\sum_{s=\lfloor \frac{T-1}{2} \rfloor + 1}^{T-1} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \sum_{t=1}^{s-1} \frac{1}{(T-t)^2}}_{\mathcal{T}_2}.
\end{aligned}$$

Using $s \leq \frac{T-1}{2}$, so that $T - t \geq \frac{T}{2}$ for all $t \leq s-1$, we have

$$\begin{aligned}
\mathcal{T}_1 &\leq \sum_{s=2}^{\lfloor \frac{T-1}{2} \rfloor} \left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \frac{4s}{T^2} \\
&\leq 4M_\nu^{\frac{\nu}{1+\nu}} \frac{(1+\log(T))^\nu}{T^2} \sum_{s=2}^{\lfloor \frac{T-1}{2} \rfloor} \left(35 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta}}\right)^\nu s^{1-\nu} + 15 \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^\nu s^{1-\frac{\nu}{2}}\right) \\
&\leq 4M_\nu^{\frac{\nu}{1+\nu}} \frac{(1+\log(T))^\nu}{T^2} \left(35 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{2\nu}}{\tilde{\eta}^\nu}\right) T^{2-\nu} + 15 \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^\nu T^{2-\frac{\nu}{2}}\right) \\
&\leq 200 M_\nu^{\frac{\nu}{1+\nu}} (1+\log(T))^\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{2\nu}}{\tilde{\eta}^\nu T^\nu}\right),
\end{aligned}$$

where we use the induction hypothesis together with $(a+b)^p \leq a^p + b^p$ for $p = \frac{\nu}{1+\nu} \in [0, \frac{1}{2}]$ in the second inequality, $\sum_{s=2}^{\lfloor (T-1)/2 \rfloor} s^\alpha \leq T^{1+\alpha}$ for $\alpha \in [0,1]$ in the third inequality, and $\tilde{\eta} \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$, i.e., $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^\nu T^{-\nu/2} \leq \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T}\right)^\nu$, in the last inequality. Using $s \geq \lfloor \frac{T-1}{2} \rfloor + 1$, i.e., $s \geq \frac{T}{2}$, and $T \geq 2$, we have

$$\begin{aligned}
\mathcal{T}_2 &\leq (1+\log(T))^\nu M_\nu^{\frac{\nu}{1+\nu}} \sum_{s=\lfloor \frac{T-1}{2} \rfloor}^{T-1} \left(70 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T}\right)^\nu + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\right)^{\frac{\nu}{2}}\right) \sum_{t=1}^{s-1} \frac{1}{(T-t)^2} \\
&\leq (1+\log(T))^\nu M_\nu^{\frac{\nu}{1+\nu}} \sum_{s=\lfloor \frac{T-1}{2} \rfloor}^{T-1} \left(70 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T}\right)^\nu + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\right)^{\frac{\nu}{2}}\right) \frac{1}{T-s} \\
&\leq (1+\log(T))^{1+\nu} M_\nu^{\frac{\nu}{1+\nu}} \left(70 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T}\right)^\nu + 30 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\right)^{\frac{\nu}{2}}\right),
\end{aligned}$$

where the first inequality uses the induction hypothesis as for $\mathcal{T}_1$ together with $s \geq \frac{T}{2}$ and extends the range of summation to $s \geq \lfloor \frac{T-1}{2} \rfloor$, the second inequality uses $\sum_{t=1}^{s-1} \frac{1}{(T-t)^2} \leq \sum_{u=T-s+1}^{\infty} \frac{1}{u^2} \leq \frac{1}{T-s}$, and the last inequality uses $\sum_{s=\lfloor (T-1)/2 \rfloor}^{T-1} \frac{1}{T-s} \leq 1 + \log(T)$. Combining the above inequalities, we get

$$\begin{aligned}
2\left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right) &\leq 6\tilde{\eta} M_\nu^{\frac{1}{1+\nu}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} \\
&\quad + (1+\log(T))^{1+\nu} M_\nu \left(860 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\tilde{\eta} T}\right)^{1+\nu} + 60 \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}\right).
\end{aligned}$$

Then, using $\tilde{\eta} \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$ and $T \ge 1$, we obtain

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \le \frac{4\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T\tilde{\eta}} M_\nu^{\frac{1}{1+\nu}} \left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right)^{\frac{\nu}{1+\nu}} + (1+\log(T))^{1+\nu} M_\nu \left(430\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{\tilde{\eta}T}\right)^{1+\nu} + 30\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}\right).$$

From Lemma 5, we obtain

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \le (1+\log(T))^{1+\nu} M_\nu \left(876\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{\tilde{\eta}T}\right)^{1+\nu} + 60\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}\right).$$

This implies Eq. (10) for $t = T$, which completes the induction. Finally, substituting $\tilde{\eta} = \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$ and $t = T$ into Eq. (10) proves Theorem 1 with the absolute constant 1400; we have not attempted to optimize this constant. □

D Proof of Theorem 2

Notation. Throughout this section, $\tilde{\eta}_t = (1 - \frac{t}{T+1})\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$ for $t \ge 1$, and we also set $\tilde{\eta}_0 := \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$. We use the following notation:

$$\begin{aligned} f^\star &:= f(\boldsymbol{x}^\star), \\ \Phi(\boldsymbol{x}) &:= (f(\boldsymbol{x}) - f^\star)^{\frac{1}{1+\nu}}, \\ J &:= \left\lceil \log_2\left(\frac{T+1}{2\sqrt{T}}\right)\right\rceil, \\ \mathcal{I}_j &:= \left\{ t \in \{1,\dots,T\} \mid 2^{-(j+1)} < 1 - \frac{t}{T+1} \le 2^{-j} \right\} \quad \text{for all } j \in \{0,\dots,J\}. \end{aligned}$$

By the definition of J, we have $\frac{\sqrt{T}}{T+1} < 2^{-J} \le \frac{2\sqrt{T}}{T+1}$. The blocks $\mathcal{I}_0,\dots,\mathcal{I}_J$ are consecutive intervals of integers: $\mathcal{I}_0 = \{1,\dots,\lfloor T/2\rfloor\}$, every element of $\mathcal{I}_{j-1}$ is smaller than every element of $\mathcal{I}_j$, and, for $T \ge 16$, $\max\mathcal{I}_J \le T-2$ and $|\mathcal{I}_J| \ge 1$, because $(T+1)2^{-(J+1)} > \frac{\sqrt{T}}{2} \ge 2$. The iterations $t > \max\mathcal{I}_J$ that are not covered by the blocks satisfy $T+1-t \le (T+1)2^{-(J+1)} \le \sqrt{T}$ and $\tilde{\eta}_t \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{T+1}$.

Lemma 6. *Suppose that Assumptions 1 and 2 hold. If $f(\boldsymbol{x}) \ge f(\boldsymbol{y})$ and $\nabla f(\boldsymbol{x}) \ne 0$, it holds that*

$$\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y}) \le 2M_\nu^{\frac{1}{1+\nu}} \left\langle \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}, \boldsymbol{x}-\boldsymbol{y}\right\rangle,$$

where $\Phi(\boldsymbol{x}) := (f(\boldsymbol{x}) - f^\star)^{\frac{1}{1+\nu}}$. In particular, when $\boldsymbol{y} = \boldsymbol{x}^\star$ we have

$$\Phi(\boldsymbol{x}) \le 2M_\nu^{\frac{1}{1+\nu}} \left\langle \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}, \boldsymbol{x}-\boldsymbol{x}^\star\right\rangle.$$

Proof. Using $f(\boldsymbol{x}) \ge f(\boldsymbol{y})$, we have

$$\begin{aligned} (f(\boldsymbol{x}) - f^\star)^{\frac{1}{1+\nu}} - (f(\boldsymbol{y}) - f^\star)^{\frac{1}{1+\nu}} &\le (f(\boldsymbol{x}) - f^\star)^{\frac{-\nu}{1+\nu}} \left(f(\boldsymbol{x}) - f(\boldsymbol{y})\right) \\ &\le (f(\boldsymbol{x}) - f^\star)^{\frac{-\nu}{1+\nu}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle \\ &\le 2M_\nu^{\frac{1}{1+\nu}} \left\langle \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}, \boldsymbol{x}-\boldsymbol{y}\right\rangle, \end{aligned}$$

where the first inequality uses $a^p - b^p \le a^{p-1}(a-b)$ for $a \ge b \ge 0$ and $p = \frac{1}{1+\nu} \in [\frac{1}{2}, 1]$, the second inequality uses Assumption 1, and the last inequality uses $\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle \ge f(\boldsymbol{x}) - f(\boldsymbol{y}) \ge 0$ and Lemma 4 in the form $(f(\boldsymbol{x}) - f^\star)^{-\frac{\nu}{1+\nu}} \|\nabla f(\boldsymbol{x})\| \le 2M_\nu^{\frac{1}{1+\nu}}$. □

Lemma 7. *Suppose that Assumptions 1 and 2 hold. Let $1 \leq s \leq p \leq q \leq T$. There exists $r \in \{p, \dots, q\}$ such that*

$$\Phi(\boldsymbol{x}_r) \leq \Phi(\boldsymbol{x}_s) + M_\nu^{\frac{1}{1+\nu}} \frac{\sum_{t=s}^q \tilde{\eta}_t^2}{\sum_{t=p}^q \tilde{\eta}_t},$$

where $\Phi(\boldsymbol{x}) \coloneqq (f(\boldsymbol{x}) - f^\star)^{\frac{1}{1+\nu}}$.

Proof. Choose $s^\star \in \{s, \dots, p\}$ minimizing $f(\boldsymbol{x}_t)$ on this interval.

If some $r \in \{p, \dots, q\}$ satisfies $f(\boldsymbol{x}_r) < f(\boldsymbol{x}_{s^\star})$, the statement holds, since then $\Phi(\boldsymbol{x}_r) < \Phi(\boldsymbol{x}_{s^\star}) \leq \Phi(\boldsymbol{x}_s)$. Otherwise, $f(\boldsymbol{x}_t) \geq f(\boldsymbol{x}_{s^\star})$ for all $t \in \{p, \dots, q\}$. For any $t \in \{p, \dots, q\}$, Lemma 6 implies that

$$\Phi(\boldsymbol{x}_t) - \Phi(\boldsymbol{x}_{s^\star}) \leq 2M_\nu^{\frac{1}{1+\nu}} \left\langle \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \boldsymbol{x}_t - \boldsymbol{x}_{s^\star} \right\rangle.$$

The update rule of normalized gradient descent gives

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{s^\star}\|^2 = \|\boldsymbol{x}_t - \boldsymbol{x}_{s^\star}\|^2 - 2\tilde{\eta}_t \left\langle \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \boldsymbol{x}_t - \boldsymbol{x}_{s^\star} \right\rangle + \tilde{\eta}_t^2.$$

Summing this identity from $t = s^\star$ to q and dropping the nonnegative term $\|\boldsymbol{x}_{q+1} - \boldsymbol{x}_{s^\star}\|^2$ gives $2\sum_{t=s^\star}^q \tilde{\eta}_t \langle \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \boldsymbol{x}_t - \boldsymbol{x}_{s^\star} \rangle \leq \sum_{t=s^\star}^q \tilde{\eta}_t^2$. Moreover, for $t \in \{s^\star, \dots, p-1\}$ we have $f(\boldsymbol{x}_t) \geq f(\boldsymbol{x}_{s^\star})$ by the choice of $s^\star$, so Assumption 1 gives $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_{s^\star} \rangle \geq f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{s^\star}) \geq 0$; hence, the terms with $t < p$ in the sum on the left-hand side are nonnegative and can be dropped. Combining these facts with the previous inequality yields

$$\begin{aligned}
\sum_{t=p}^q \tilde{\eta}_t \left(\Phi(\boldsymbol{x}_t) - \Phi(\boldsymbol{x}_{s^\star})\right) &\leq 2M_\nu^{\frac{1}{1+\nu}} \sum_{t=s^\star}^q \tilde{\eta}_t \left\langle \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \boldsymbol{x}_t - \boldsymbol{x}_{s^\star} \right\rangle \\
&\leq M_\nu^{\frac{1}{1+\nu}} \sum_{t=s^\star}^q \tilde{\eta}_t^2 \\
&\leq M_\nu^{\frac{1}{1+\nu}} \sum_{t=s}^q \tilde{\eta}_t^2,
\end{aligned}$$

where we use $s \leq s^\star \leq p$. Choosing $r \in \{p, \dots, q\}$ minimizing $\Phi(\boldsymbol{x}_t)$ and using $\big(\Phi(\boldsymbol{x}_r) - \Phi(\boldsymbol{x}_{s^\star})\big) \sum_{t=p}^q \tilde{\eta}_t \leq \sum_{t=p}^q \tilde{\eta}_t (\Phi(\boldsymbol{x}_t) - \Phi(\boldsymbol{x}_{s^\star}))$ yields

$$\Phi(\boldsymbol{x}_r) - \Phi(\boldsymbol{x}_{s^\star}) \leq M_\nu^{\frac{1}{1+\nu}} \frac{\sum_{t=s}^q \tilde{\eta}_t^2}{\sum_{t=p}^q \tilde{\eta}_t}.$$

Since $\Phi(\boldsymbol{x}_{s^\star}) \leq \Phi(\boldsymbol{x}_s)$, this concludes the proof. □

Lemma 8. *Suppose that Assumptions 1 and 2 hold, $T \geq 16$, and $\tilde{\eta}_t = (1 - \frac{t}{T+1}) \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$. For any $j \in \{0, \dots, J\}$, there exists $s_j \in \mathcal{I}_j$ such that*

$$f(\boldsymbol{x}_{s_j}) - f^\star \leq 23104 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}}.$$

Proof. For $t \in \mathcal{I}_j$, we have, by the definition of $\mathcal{I}_j$

$$\frac{1}{2^{j+1}} \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} < \tilde{\eta}_t \leq \frac{1}{2^j} \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}, \tag{15}$$

$$\frac{T+1}{2^{j+1}} - 1 \leq |\mathcal{I}_j| \leq \frac{T+1}{2^{j+1}} + 1. \tag{16}$$

From Lemma 6, we obtain

$$\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t\Phi(\boldsymbol{x}_t)\le 2M_\nu^{\frac{1}{1+\nu}}\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t\left\langle\frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|},\boldsymbol{x}_t-\boldsymbol{x}^\star\right\rangle$$
$$\le 2M_\nu^{\frac{1}{1+\nu}}\left(\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2+\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t^2\right),$$

where we use the following in the last inequality:

$$\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^2\le\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2-2\tilde{\eta}_t\left\langle\frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|},\boldsymbol{x}_t-\boldsymbol{x}^\star\right\rangle+\tilde{\eta}_t^2.$$

Since $\mathcal{I}_0=\{1,\dots,\lfloor T/2\rfloor\}$, a direct computation gives

$$\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t=\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\sum_{t=1}^{\lfloor T/2\rfloor}\left(1-\frac{t}{T+1}\right)\ge\frac{3(T-1)}{8}\cdot\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}\ge\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|\frac{\sqrt{T+1}}{4},$$

where the last inequality holds for $T\ge 4$ (it is equivalent to $5T^2-22T+9\ge 0$), and Eqs. (15) and (16) give

$$\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t^2\le\left(\frac{T+1}{2}+1\right)\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\le 2\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2.$$

Since $\min_{t\in\mathcal{I}_0}\Phi(\boldsymbol{x}_t)\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t\le\sum_{t\in\mathcal{I}_0}\tilde{\eta}_t\Phi(\boldsymbol{x}_t)\le 6M_\nu^{\frac{1}{1+\nu}}\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2$, we conclude that there exists $s_0\in\mathcal{I}_0$ such that

$$\Phi(\boldsymbol{x}_{s_0})\le 24M_\nu^{\frac{1}{1+\nu}}\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}.\tag{17}$$

Next, we discuss the case when $j\ge 1$. Choose s_j as follows for any j:

$$s_j\coloneqq\operatorname*{arg\,min}_{t\in\mathcal{I}_j}\Phi(\boldsymbol{x}_t).$$

From Lemma 7, it holds for any $j\ge 1$

$$\Phi(\boldsymbol{x}_{s_j})\le\Phi(\boldsymbol{x}_{s_{j-1}})+M_\nu^{\frac{1}{1+\nu}}\frac{\sum_{t\in\mathcal{I}_{j-1}\cup\mathcal{I}_j}\tilde{\eta}_t^2}{\sum_{t\in\mathcal{I}_j}\tilde{\eta}_t}.$$

Using Eqs. (15) and (16), $T\ge 16$, and $2^j\le 2^J<\frac{T+1}{\sqrt{T}}$, we obtain

$$\sum_{t\in\mathcal{I}_{j-1}\cup\mathcal{I}_j}\tilde{\eta}_t^2\le\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\left[\left(\frac{T+1}{2^j}+1\right)\frac{1}{4^{j-1}}+\left(\frac{T+1}{2^{j+1}}+1\right)\frac{1}{4^j}\right]$$
$$=\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{8^j}\left[\frac{9(T+1)}{2T}+\frac{5\cdot 2^j}{T}\right]\le\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{8^j}\cdot\frac{391}{64}\le\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2\frac{1}{2^{3j-4}},$$
$$\sum_{t\in\mathcal{I}_j}\tilde{\eta}_t\ge\left(\frac{T+1}{2^{j+1}}-1\right)\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{2^{j+1}\sqrt{T}}\ge\frac{(T+1)\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{2^{2j+3}\sqrt{T}}\ge\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|\frac{\sqrt{T+1}}{2^{2j+3}},$$

where, in the second line, we use $\frac{T+1}{2^{j+1}}-1\ge\frac{T+1}{2^{j+2}}$, which holds because $2^{j+2}\le 2^{J+2}<\frac{4(T+1)}{\sqrt{T}}\le T+1$ for $T\ge 16$. Therefore, we get

$$\Phi(\boldsymbol{x}_{s_j})\le\Phi(\boldsymbol{x}_{s_{j-1}})+M_\nu^{\frac{1}{1+\nu}}\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T+1}}\frac{1}{2^{j-7}}.$$

Unrolling this recursion, using Eq. (17) and $\sum_{i=1}^j 2^{7-i}\le 2^7$, we obtain $\Phi(\boldsymbol{x}_{s_j})\le(24+128)M_\nu^{\frac{1}{1+\nu}}\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}=152M_\nu^{\frac{1}{1+\nu}}\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}}$ for all $j\in\{0,\dots,J\}$. Raising both sides to the power $1+\nu\le 2$ gives the statement. □

Lemma 9. *Suppose that Assumptions 1 and 2 hold, $T \geq 16$, and $\tilde{\eta}_t = (1 - \frac{t}{T+1}) \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$. We define*

$$\mathcal{E}(r) \coloneqq \sum_{t=r}^{T} \tilde{\eta}_t \|\nabla f(\boldsymbol{x}_t)\|,$$
$$s_j \coloneqq \operatorname*{arg\,min}_{t \in \mathcal{I}_j} f(\boldsymbol{x}_t).$$

We have

$$\begin{cases} \mathcal{E}(1) & \leq 2M_\nu \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{1+\nu} (T+1)^{\frac{1-\nu}{2}}, \\ \mathcal{E}(s_{j-1}) & \leq M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} \left(23104 + \frac{T+1}{2^{(j-1)(2+\nu)}} \right) \quad \textit{for all } j \in \{1, \dots, J\}, \\ \mathcal{E}(s_J) & \leq 23112 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}}. \end{cases}$$

Proof. Note that s_j is the choice made in the proof of Lemma 8, so the bound of that lemma applies to s_j for every $j \in \{0, \dots, J\}$. From Lemma 3 (using $\frac{M_\nu}{1+\nu} \leq M_\nu$), we have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + M_\nu \tilde{\eta}_t^{1+\nu} \\ &\leq f(\boldsymbol{x}_t) - \tilde{\eta}_t \|\nabla f(\boldsymbol{x}_t)\| + M_\nu \tilde{\eta}_t^{1+\nu}. \end{aligned}$$

Summing up the above inequalities from $t = s$ to T yields

$$\mathcal{E}(s) = \sum_{t=s}^{T} \tilde{\eta}_t \|\nabla f(\boldsymbol{x}_t)\| \leq f(\boldsymbol{x}_s) - f^\star + M_\nu \sum_{t=s}^{T} \tilde{\eta}_t^{1+\nu}.$$

We first consider the case when $s = 1$. Assumption 2 and Lemma 3 imply that

$$\begin{aligned} \mathcal{E}(1) &\leq M_\nu \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{1+\nu} + M_\nu \sum_{t=1}^{T} \tilde{\eta}_t^{1+\nu} \\ &\leq M_\nu \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{1+\nu} + M_\nu \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{1+\nu} (T+1)^{\frac{1-\nu}{2}} \\ &\leq 2M_\nu \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{1+\nu} (T+1)^{\frac{1-\nu}{2}}. \end{aligned}$$

Next, we consider the case when $j \in \{1, \dots, J\}$. From Lemma 8, we obtain

$$\begin{aligned} \mathcal{E}(s_{j-1}) &\leq 23104 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} + M_\nu \sum_{t=s_{j-1}}^{T} \tilde{\eta}_t^{1+\nu} \\ &\leq M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} \left(23104 + \frac{T+1}{2^{(j-1)(2+\nu)}} \right), \end{aligned}$$

where we use that $\tilde{\eta}_t \leq 2^{-(j-1)} \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ for all $t \geq s_{j-1}$ and that the number of such t is $T + 1 - s_{j-1} \leq (T+1)2^{-(j-1)}$, both by the definition of $s_{j-1} \in \mathcal{I}_{j-1}$.

Finally, we have

$$\begin{aligned} \mathcal{E}(s_J) &\leq 23104 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} + M_\nu \sum_{t=s_J}^{T} \tilde{\eta}_t^{1+\nu} \\ &\leq 23112 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}}, \end{aligned}$$

where we use $T - s_J + 1 \leq (T+1)2^{-J} \leq 2\sqrt{T}$ and $\tilde{\eta}_t \leq 2^{-J} \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} \leq \frac{2\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{T+1}$ for $t \geq s_J$, so that $\sum_{t=s_J}^{T} \tilde{\eta}_t^{1+\nu} \leq 2\sqrt{T} \big(\frac{2\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{T+1} \big)^{1+\nu} \leq 8 \big(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \big)^{\frac{1+\nu}{2}}$, in the last inequality. □

Lemma 10. *Suppose that Assumptions 1 and 2 hold. Normalized gradient descent with* $\tilde{\eta}_t = (1 - \frac{t}{T+1})\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ *satisfies*

$$\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \leq 2\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2,$$

for any $t \in \{1, \ldots, T+1\}$.

Proof. From the update rule of normalized gradient descent, we have

$$\begin{aligned}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 &\leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - 2\tilde{\eta}_t \left\langle \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}, \boldsymbol{x}_t - \boldsymbol{x}^\star \right\rangle + \tilde{\eta}_t^2 \\ &\leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \tilde{\eta}_t^2,\end{aligned}$$

where we use Assumption 1 in the second inequality. Recursively applying the above inequality, we obtain

$$\begin{aligned}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 &\leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{s=1}^{t} \tilde{\eta}_s^2 \\ &\leq 2\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2,\end{aligned}$$

where the last inequality uses $\tilde{\eta}_s \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}}$ and $t \leq T$. □

Proof of Theorem 2. We show that

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq 277384\, M_\nu \Big(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\Big)^{\frac{1+\nu}{2}},$$

for every $T \geq 1$; we have not attempted to optimize this constant. If $T < 16$, then Lemma 3 with $\boldsymbol{x} = \boldsymbol{x}^\star$ and Lemma 10 give $f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{M_\nu}{1+\nu}\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^{1+\nu} \leq M_\nu\big(2\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2\big)^{\frac{1+\nu}{2}} \leq 32 M_\nu\big(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T}\big)^{\frac{1+\nu}{2}}$, where we use $T < 16$ in the last step. We now consider the case $T \geq 16$.

Let $\{v_t\}_{t=0}^T$ be defined as follows:

$$\begin{aligned}v_T &\coloneqq \frac{1}{\eta_T}, \\ v_t &\coloneqq \frac{1}{\sum_{s=t+1}^T \eta_s}, \quad (\text{for all } t = 0, 1, \ldots, T-1).\end{aligned}$$

As in the proof of Lemma 2, $\{v_t\}_{t=0}^T$ is positive and nondecreasing,

$$\eta_T v_{T-1} = \eta_T v_T = 1,$$

and for any $t = 1, 2, \ldots, T-1$, we have

$$\eta_t v_t - \left(\sum_{s=t}^T \eta_s\right)(v_t - v_{t-1}) = -\left(\sum_{s=t+1}^T \eta_s\right) v_t + \left(\sum_{s=t}^T \eta_s\right) v_{t-1} = 0.$$

Thus, using Lemma 1, we obtain

$$2\left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right) \leq v_0 \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^T v_t \tilde{\eta}_t^2.$$

From the Cauchy–Schwarz inequality, it holds that for any $t \in \{0, 1, \ldots, T-1\}$,

$$\left(\sum_{s=t+1}^T \frac{\tilde{\eta}_s}{\|\nabla f(\boldsymbol{x}_s)\|}\right)\left(\sum_{s=t+1}^T \tilde{\eta}_s \|\nabla f(\boldsymbol{x}_s)\|\right) \geq \left(\sum_{s=t+1}^T \tilde{\eta}_s\right)^2 = \tilde{\eta}_t^2 \frac{(T-t)^2}{4}. \tag{18}$$

This implies that

$$v_t \le \frac{4\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|}{\tilde{\eta}_t^2(T-t)^2}.$$

In particular, we get

$$v_0 \le \frac{4\sum_{s=1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|}{\tilde{\eta}_0^2T^2} = \frac{4\mathcal{E}(1)}{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2T} \le 16M_\nu\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{-1+\nu}\left(\frac{1}{T}\right)^{\frac{1+\nu}{2}},$$

where we use Lemma 9 and $(T+1)^{\frac{1-\nu}{2}} \le 2T^{\frac{1-\nu}{2}}$. Since $\mathcal{I}_0,\dots,\mathcal{I}_J$ partition $\{1,\dots,\max\mathcal{I}_J\}$ and $\max\mathcal{I}_J \le T-2$, we can split the sum over t as follows:

$$\begin{aligned}
&2\left(f(\boldsymbol{x}_T)-f(\boldsymbol{x}^\star)\right)\\
&\le 16M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}} + v_T\tilde{\eta}_T^2 + \sum_{t=1}^{T-1}\frac{4}{(T-t)^2}\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|\\
&= 16M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}} + v_T\tilde{\eta}_T^2\\
&\quad + \sum_{j=0}^{J}\sum_{t\in\mathcal{I}_j\wedge t<T}\frac{4}{(T-t)^2}\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\| + \sum_{t=1+\max\mathcal{I}_J}^{T-1}\frac{4}{(T-t)^2}\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|.
\end{aligned}$$

Moreover, we have, for every $j\in\{0,\dots,J\}$,

$$\sum_{t\in\mathcal{I}_j\wedge t<T}\frac{1}{(T-t)^2} \le \frac{2^{j+3}}{T+1}. \tag{19}$$

Indeed, for $t\in\mathcal{I}_j$ we have $T+1-t > (T+1)2^{-(j+1)}$ and hence, since $T+1-t$ is an integer, $T-t \ge \lfloor (T+1)2^{-(j+1)}\rfloor \ge (T+1)2^{-(j+2)}$, where the last inequality uses $(T+1)2^{-(j+1)} \ge (T+1)2^{-(J+1)} \ge 2$; moreover, $\sum_{1\ge u\ge m}\frac{1}{u^2} \le \frac{2}{m}$ for any integer $m\ge 1$.

Using Lemma 9, $\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\| = \mathcal{E}(t+1) \le \mathcal{E}(1)$, and Eq. (19), we get

$$\begin{aligned}
\sum_{t\in\mathcal{I}_0}\frac{4}{(T-t)^2}\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\| &\le 8M_\nu\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{1+\nu}(T+1)^{\frac{1-\nu}{2}}\sum_{t\in\mathcal{I}_0}\frac{1}{(T-t)^2}\\
&\le 64M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T+1}\right)^{\frac{1+\nu}{2}} \le 64M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}.
\end{aligned}$$

Similarly, since $s_{j-1}\in\mathcal{I}_{j-1}$ implies $s_{j-1}\le t$ for every $t\in\mathcal{I}_j$, we have $\mathcal{E}(t+1)\le\mathcal{E}(s_{j-1})$ for $t\in\mathcal{I}_j$, and hence

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{t\in\mathcal{I}_j\wedge t<T}\frac{4}{(T-t)^2}\sum_{s=t+1}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|\\
&\le \sum_{j=1}^{J}\sum_{t\in\mathcal{I}_j\wedge t<T}\frac{4}{(T-t)^2}\sum_{s=s_{j-1}}^{T}\tilde{\eta}_s\|\nabla f(\boldsymbol{x}_s)\|\\
&\le M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}\sum_{j=1}^{J}\sum_{t\in\mathcal{I}_j\wedge t<T}\frac{4}{(T-t)^2}\left(23104+\frac{T+1}{2^{(j-1)(2+\nu)}}\right)\\
&\le M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}}\sum_{j=1}^{J}\left(92416\frac{2^{j+3}}{T+1}+\frac{64}{2^{(j-1)(1+\nu)}}\right)\\
&\le 369792M_\nu\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}\right)^{\frac{1+\nu}{2}},
\end{aligned}$$

where we use Lemma 9 and Eq. (19) for the second and third inequalities, and $\sum_{j=1}^{J} 2^{j+3} \leq 2^{J+4} < \frac{16(T+1)}{\sqrt{T}} \leq 4(T+1)$ together with $\sum_{j=1}^{J} 2^{-(j-1)(1+\nu)} \leq 2$ for the last inequality. For the remaining iterations $t > \max \mathcal{I}_J \geq s_J$, we similarly have

$$\begin{aligned}\sum_{t=1+\max \mathcal{I}_J}^{T-1} \frac{4}{(T-t)^2} \sum_{s=t+1}^{T} \tilde{\eta}_s \|\nabla f(\boldsymbol{x}_s)\| &\leq \sum_{t=1+\max \mathcal{I}_J}^{T-1} \frac{4}{(T-t)^2} \sum_{s=s_J}^{T} \tilde{\eta}_s \|\nabla f(\boldsymbol{x}_s)\| \\ &\leq 7 \sum_{s=s_J}^{T} \tilde{\eta}_s \|\nabla f(\boldsymbol{x}_s)\| \\ &\leq 161784 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}},\end{aligned}$$

where we use $4 \sum_{u=1}^{\infty} \frac{1}{u^2} = \frac{2\pi^2}{3} \leq 7$ in the second inequality and Lemma 9 in the last inequality. Finally, we also have

$$v_T \tilde{\eta}_T^2 = \tilde{\eta}_T \|\nabla f(\boldsymbol{x}_T)\| \leq \sum_{t=s_J}^{T} \tilde{\eta}_t \|\nabla f(\boldsymbol{x}_t)\| \leq 23112 M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}},$$

where we use Lemma 9 in the last inequality. Combining the above inequalities, we obtain

$$\begin{aligned}2\left(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star)\right) &\leq (16 + 64 + 369792 + 161784 + 23112)\, M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}} \\ &= 554768\, M_\nu \left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} \right)^{\frac{1+\nu}{2}},\end{aligned}$$

which is the desired result. □