STOCHASTIC NON-SMOOTH CONVEX OPTIMIZATION WITH UNBOUNDED GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Much of the existing theory on first-order non-smooth optimization is built on a restrictive assumption that the gradients of the objective function are uniformly bounded. We introduce a much more realistic class of generalized Lipschitz functions, where the gradient norms are bounded by an affine function of the optimality gap. We then ask a natural question: what algorithm achieves the best global convergence rates for solving convex stochastic generalized Lipschitz optimization problems? To address this, we develop a new convergence analysis for several existing algorithms and find that AdamW with clipped updates, provably outperforms other popular stochastic optimization methods, such as SGD and AdaGrad. Moreover, our analysis establishes the critical role of AdamW's exponentially weighted gradient accumulation, as opposed to simple averaging. We further show that clipped AdamW is universal and achieves improved rates under the popular generalized smoothness assumption, analyze the convergence of clipped AdamW with diagonal and matrix preconditioners, and extend our results to the quasar-convex setting.

1 INTRODUCTION

Motivation. Stochastic non-smooth first-order optimization methods play a central role in modern machine learning, particularly in training deep neural networks with non-smooth components such as ReLU activations and max-pooling layers (Goodfellow et al., 2016; He et al., 2015; Glorot et al., 2011). The convergence properties of these methods are typically analyzed under the assumption that the objective function is globally Lipschitz, or equivalently, that its gradients (or their appropriate generalizations) are uniformly bounded (Nesterov et al., 2018; Shor, 2012; Zhang et al., 2020). Although this assumption is mathematically convenient and leads to clean worst-case convergence guarantees, it also imposes significant limitations. In particular, global Lipschitz continuity may fail even for the training loss of simple two-layer ReLU networks. Moreover, theory based on bounded gradients provides limited insight into the empirical success of Adam-type methods. Despite extensive efforts to develop better optimizers for deep learning, Adam and AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017) remain widely used and highly competitive, often outperforming other adaptive and non-adaptive methods such as AdaGrad (Duchi et al., 2011), SGD (Robbins & Monro, 1951), and Signum (Bernstein et al., 2018). Yet, under the bounded gradients assumption, simple SGD or its momentum-based variants already attain optimal convergence rates in both convex (Nemirovski et al., 2009) and non-convex (Cutkosky et al., 2023) settings. Therefore, even in the fundamental setting of non-smooth convex optimization, which is the focus of this paper, bounded-gradient theory leaves a clear gap in the understanding of the advantages of Adam-type algorithms over other first-order methods. This motivates the search for a suitable relaxation or generalization of the bounded-gradient assumption, together with an appropriate convergence theory capable of closing this gap.

Summary of contributions. Motivated by the above background, we present the following main contributions.

- **(i)** In Section 2, we introduce the class of $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitz, or generalized Lipschitz, functions and derive their basic properties.
- **(ii)** In Section 3, we analyze the convergence of several deterministic and stochastic first-order methods for convex generalized Lipschitz functions. The results are summarized in Table 1

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.

and highlight the limitations of SGD and AdaGrad for stochastic generalized Lipschitz problems.

(iii) In Section 4, we show that scalar-stepsize AdamW with clipped updates and exponential decay significantly improves upon the convergence guarantees for SGD and AdaGrad obtained in Section 3 for stochastic convex generalized Lipschitz problems.

(iv) In Section A, we establish lower complexity bounds for SGD and AdaGrad on stochastic convex generalized Lipschitz problems, showing that the improvement achieved by scalar-stepsize AdamW with clipped updates is genuine rather than an artifact of the analysis.

(v) In Section B, we show that clipped scalar-stepsize AdamW is also effective under the generalized Hölder smoothness assumption, which interpolates between $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitzness and the popular $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness assumption (Zhang et al., 2019).

(vi) In Section C, we extend our results to the quasar-convex setting (Hinder et al., 2020), which relaxes the standard convexity assumption.

(vii) In Section D, we discuss that our results are not specific to scalar stepsizes and can be extended to diagonal and matrix preconditioners.

Notations. In this paper we use the following notation: $\|\cdot\|_2$ denotes the standard Euclidean norm, induced by the standard inner product $\langle\cdot,\cdot\rangle$; $\|\cdot\|_\infty$ denotes the infinity vector norm; $\|\cdot\|_{\text{op}}$ denotes the matrix spectral norm; for $x \in \mathbb{R}^d$, $\text{clip}_2(x)$ denotes the Euclidean norm clipping with unit radius, i.e., $\text{clip}_2(x) = x \cdot \min\{1, 1/\|x\|_2\}$, and $\text{clip}_\infty(x)$ denotes the coordinate-wise clipping with unit radius; $\odot$ and $\oslash$ denote elementwise multiplication and division, respectively; $\mathcal{O}[\cdot]$ and $\Omega[\cdot]$ hide universal constant multiplicative factors; $\tilde{\mathcal{O}}[\cdot]$ hides universal constant and polylogarithmic multiplicative factors, depending on the assumption constants and target accuracy; $\mathbb{N}$ and $\mathbb{N}_0$ denote the sets of positive and non-negative integers, respectively. We also use the conventions $1/0 = +\infty$ and $0/0 = 0$; whenever an expression defines a function $\psi(t)$ for $t > 0$ that admits a continuous extension to $t = 0$, we set $\psi(0) = \lim_{t\to+0} \psi(t)$.

2 Generalized Lipschitz Optimization

Throughout this paper, we focus on the problem of minimizing a locally Lipschitz, possibly non-smooth, objective function $f(x)\colon \mathbb{R}^d \to \mathbb{R}$. We assume that at least one global solution $x^* \in \mathbb{R}^d$ exists and denote the corresponding optimal function value as $f^* = f(x^*)$. We also assume that $\|x^*\|_2 \le \mathcal{R}$ for some positive constant $\mathcal{R} > 0$.

2.1 Generalized Lipschitz Functions

As discussed in Section 1, assuming global Lipschitz continuity is restrictive. Therefore, we introduce a relaxed condition, which we refer to as $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitzness or generalized Lipschitzness. In the case where the objective function $f(x)$ is differentiable, this condition can be stated as follows:

$$\|\nabla f(x)\|_2 \le \mathcal{M}_0 + \mathcal{M}_1[f(x) - f^*] \quad \text{for all} \quad x \in \mathbb{R}^d, \tag{1}$$

which clearly generalizes the standard bounded gradients assumption. In the case where the function $f(x)$ is not differentiable, we provide even more general Assumption 1.

Assumption 1. *The function $f(x)$ is $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitz. That is, the following inequality holds:*

$$\limsup_{u\to 0} \frac{1}{\|u\|_2}|f(x+u) - f(x)| \le \mathcal{M}_0 + \mathcal{M}_1[f(x) - f^*] \quad \textit{for all} \quad x \in \mathbb{R}^d. \tag{2}$$

Motivation. Our main theoretical motivation for condition (1) comes from its parallel with the following $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness condition, also known as generalized smoothness, introduced by Zhang et al. (2019) for twice continuously differentiable functions:

$$\|\nabla^2 f(x)\|_{\text{op}} \le \mathcal{L}_0 + \mathcal{L}_1\|\nabla f(x)\|_2 \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{3}$$

This condition was introduced to relax the standard smoothness assumption commonly used in smooth first-order optimization (Nesterov et al., 2018), namely, the global Lipschitz continuity of the gradient. Observe that condition (3) bounds the second derivatives by an affine function of the first derivatives. This suggests an analogous condition for non-smooth first-order optimization: reducing the order of the derivatives by one and recalling that the zeroth derivative is the function itself leads precisely

to condition (1). Additionally, Lemma 2.2 of Gorbunov et al. (2024) shows that the generalized smoothness implies the more general bound

$$\|\nabla f(x)\|_2 \leq \mathcal{O}\big[\sqrt{\mathcal{L}_0[f(x) - f^*]} + \mathcal{L}_1[f(x) - f^*]\big]. \tag{4}$$

On the other hand, standard global $\mathcal{L}_0$-smoothness and $\mathcal{M}_0$-Lipschitzness imply, respectively, $\|\nabla f(x)\|_2 \leq \mathcal{O}\big[\sqrt{\mathcal{L}_0[f(x) - f^*]}\big]$ and $\|\nabla f(x)\|_2 \leq \mathcal{M}_0$. Thus, replacing the "$\mathcal{L}_0$-part" in eq. (4) with its "$\mathcal{M}_0$-analogue" leads to our condition (1). This observation also motivates the generalized Hölder smoothness condition studied in Section B, which interpolates between conditions (1) and (3).

Basic examples. To show formally that Assumption 1 strictly generalizes the standard global Lipschitz continuity assumption, we provide two simple examples. Let $p > 1$, $A \in \mathbb{R}^{n\times d}$, and $b \in \mathbb{R}^n$. Then, the function $f(x) = \|Ax - b\|_\infty^p$ is $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitz with arbitrary $\mathcal{M}_1 > 0$ and $\mathcal{M}_0 = \|A\|_{\text{op}}^p (\frac{p-1}{\mathcal{M}_1})^{p-1} + \mathcal{M}_1 f^*$, and the function $f(x) = \exp(\|Ax - b\|_\infty)$ is $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitz with $\mathcal{M}_1 = \|A\|_{\text{op}}$ and $\mathcal{M}_0 = \mathcal{M}_1 f^*$.

2.2 Basic Properties of Generalized Lipschitz Functions

In this section, we derive the basic properties of generalized Lipschitz functions. We start with the equivalent characterization of $(\mathcal{M}_0, \mathcal{M}_1)$-Lipschitz functions in the following Lemma 1.

Lemma 1 (↓). *Assumption 1 holds if and only if the following inequality holds for all $x, x' \in \mathbb{R}^d$:*

$$\begin{aligned} |f(x') - f(x)| &\leq ([\mathcal{M}_0/\mathcal{M}_1] + [f(x) - f^*])[\exp(\mathcal{M}_1\|x' - x\|_2) - 1] \\ &\leq (\mathcal{M}_0 + \mathcal{M}_1[f(x) - f^*]) \exp(\mathcal{M}_1\|x' - x\|_2)\|x' - x\|_2 \end{aligned} \tag{5}$$

In the case $\mathcal{M}_1 = 0$, the inequality in eq. (5) matches the standard definition of Lipschitz functions, where the function value difference $|f(x') - f(x)|$ is bounded by a multiple of the distance between the arguments $\|x' - x\|_2$. However, in the general case $\mathcal{M}_1 > 0$, the bound includes an additional exponential factor $\exp(\mathcal{M}_1\|x' - x\|_2)$. This factor cannot be removed in general, which can be illustrated by the basic examples mentioned above.

Lemma 1 also raises the question of whether the generalized Lipschitzness condition offers any advantage over the standard global Lipschitz continuity. Indeed, one could consider minimizing the objective function $f(x)$ over the ball $Q_\mathcal{R} = \{x : \|x\|_2 \leq \mathcal{R}\}$ of radius $\mathcal{R}$ centered at the origin, which is equivalent to the original problem due to $x^* \in Q_\mathcal{R}$. This approach is common, for instance, in adaptive methods such as AdaGrad, where projection steps on $Q_\mathcal{R}$ are used to ensure the almost sure boundedness of the iterates (Duchi et al., 2011), which is necessary for convergence guarantees. Lemma 1 implies that the function $f(x)$ is globally $\mathcal{M}$-Lipschitz on $Q_\mathcal{R}$, where

$$\mathcal{M} = \mathcal{M}_0 + \mathcal{M}_1\mathcal{F}, \qquad \mathcal{F} = \max_{x\in Q_\mathcal{R}}[f(x) - f^*]. \tag{6}$$

However, this approach implies convergence rates with extra exponential factors, since the constants $\mathcal{M}$ and $\mathcal{F}$ may scale exponentially, according to Lemma 1 and the discussion following it:

$$\mathcal{M} \leq \mathcal{M}_0 \exp(\mathcal{O}[\mathcal{R}\mathcal{M}_1]), \qquad \mathcal{F} \leq [\mathcal{M}_0/\mathcal{M}_1][\exp(\mathcal{O}[\mathcal{R}\mathcal{M}_1]) - 1]. \tag{7}$$

Consequently, such convergence rates depending on $\mathcal{M}$ and $\mathcal{F}$ can be overly pessimistic, especially when $\mathcal{R}\mathcal{M}_1 \gg 1$. In the upcoming sections, we will show that Assumption 1 can be utilized more directly, leading to fast convergence rates without any exponential factors.

3 Algorithms for Generalized Lipschitz Convex Optimization

Throughout most of this paper, we focus on the case where the objective function $f(x)$ is convex. By standard results in convex analysis (Rockafellar, 1997), the subdifferential $\partial f(x)$ is non-empty at every point $x \in \mathbb{R}^d$. We denote an arbitrary subgradient by $\nabla f(x) \in \partial f(x)$ and, with a slight abuse of language, refer to it as the gradient of the function $f(x)$ at the point $x \in \mathbb{R}^d$. Consequently, it is easy to verify that Assumption 1 holds if and only if the inequality in condition (1) holds.

Table 1: Summary of the main complexity results for solving non-stochastic and stochastic generalized Lipschitz convex optimization problems (Sections 2 and 4). Universal and logarithmic constants are omitted. The conditions $\mathcal{R}\mathcal{M}_1 \geq 1$ and $\mathcal{R}\mathcal{G}_1 \geq 1$ are assumed.

Algorithm	Theorem	Iteration Complexity
Generalized Lipschitz Problems (Assumption 1)		
GD	Corollary 1	$\exp(\mathcal{O}[\mathcal{R}\mathcal{M}_1]) \cdot [\mathcal{R}\mathcal{M}_0/\epsilon] + [\mathcal{R}\mathcal{M}_0/\epsilon]^2$
Normalized/Polyak GD	Theorem 2	$[\mathcal{R}\mathcal{M}_1]^2 + [\mathcal{R}\mathcal{M}_0/\epsilon]^2$
Stochastic Generalized Lipschitz Problems (Assumption 2)		
SGD	Theorem 3	$\exp(\mathcal{O}[\mathcal{R}\mathcal{G}_1]) \cdot [\mathcal{R}\mathcal{G}_0/\epsilon] + [\mathcal{R}\mathcal{G}_0/\epsilon]^2$
AdaGrad-Norm	Theorem 4	$\exp(\mathcal{O}[\mathcal{R}\mathcal{G}_1]) \cdot [\mathcal{R}\mathcal{G}_0/\epsilon] + [\mathcal{R}\mathcal{G}_0/\epsilon]^2$
Clipped AdamW	Corollary 4	$[\mathcal{R}\mathcal{G}_1]^4 + [\mathcal{R}\mathcal{G}_0/\epsilon]^2$

3.1 Deterministic Algorithms

In this section, we analyze the convergence of projected gradient descent (GD) under Assumption 1. This algorithm iterates according to the following update rule:

$$x_{k+1} = \mathrm{proj}_{Q_{\mathcal{R}}}(x_k - \eta_k \nabla f(x_k)), \quad x_0 = 0, \tag{8}$$

where $\eta_k > 0$ is a possibly time-varying stepsize. We first consider the constant-stepsize case in Theorem 1. Then, using the upper bound on $\mathcal{F}$ in eq. (7), we derive in Corollary 1 an explicit iteration complexity in terms of the constants $\mathcal{R}$, $\mathcal{M}_0$, $\mathcal{M}_1$, and the target accuracy ϵ.

Theorem 1 (↓). *Let Assumption 1 hold and let the iterates $\{x_k\}_{k=0}^K$ be generated by eq. (8). Let the stepsizes $\eta_k > 0$ and the number of iterations $K \in \mathbb{N}$ be defined as follows:*

$$1/\eta_k = \max\{2\mathcal{M}_1^2\mathcal{F}, \mathcal{M}_0\sqrt{K+1}/\mathcal{R}\}, \qquad K = \left\lceil [\mathcal{R}\mathcal{M}_1]^2 \cdot [\mathcal{F}/\epsilon] + [\mathcal{R}\mathcal{M}_0/\epsilon]^2 \right\rceil. \tag{9}$$

Then, the inequality $f(\overline{x}_K) - f^* \leq \mathcal{O}[\epsilon]$ holds, where $\overline{x}_K = \frac{1}{K+1}\sum_{k=0}^K x_k$.

Corollary 1. *Under the conditions of Theorem 1, it is sufficient to perform the following number of iterations to reach the precision $f(\overline{x}_K) - f^* \leq \epsilon$:*

$$K = \mathcal{O}\left[\exp(\mathcal{O}[\mathcal{R}\mathcal{M}_1]) \cdot [\mathcal{R}\mathcal{M}_0/\epsilon] + [\mathcal{R}\mathcal{M}_0/\epsilon]^2\right]. \tag{10}$$

The complexity bound in Corollary 1 improves upon the iteration complexity that could be obtained by naively using the global $\mathcal{M}$-Lipschitzness of the function $f(x)$ over $Q_{\mathcal{R}}$, as discussed in Section 2.2. It also admits the following informal interpretation: the bound in eq. (55) has an initial regime with the sublinear rate $\mathcal{O}[[\mathcal{R}\mathcal{M}_1]^2\mathcal{F}/K]$ lasting up to $K_{\text{init}} = \mathcal{O}[[\mathcal{R}\mathcal{M}_1]^2 \cdot [\mathcal{M}_1\mathcal{F}/\mathcal{M}_0]^2]$ iterations, followed by the slower rate $\mathcal{O}[\mathcal{R}\mathcal{M}_0/\sqrt{K}]$, matching the standard rate of GD for $\mathcal{M}_0$-Lipschitz convex functions. Although the latter rate cannot be improved in general due to the standard lower complexity bounds (Nesterov et al., 2018), the initial phase before reaching it can be exponentially long: by eq. (7), $K_{\text{init}} = \mathcal{O}[\exp(\mathcal{O}[\mathcal{R}\mathcal{M}_1])]$. Thus, despite the improvement over the naive global-Lipschitz reduction, the overall iteration complexity can still be poor, especially when $\mathcal{R}\mathcal{M}_1 \gg 1$.

It turns out that we can further improve the convergence rate of GD by utilizing adaptive stepsizes. In particular, we consider normalized GD (Shor, 2012) and GD with the Polyak stepsizes (Polyak, 1969), Option A and Option B in eq. (11), respectively. By Theorem 2, both methods require only $K_{\text{init}} = \mathcal{O}[[\mathcal{R}\mathcal{M}_1]^2]$ initial iterations before reaching the optimal rate $\mathcal{O}[\mathcal{R}\mathcal{M}_0/\sqrt{K+1}]$. Consequently, the overall iteration complexity of normalized GD and Polyak GD contains no exponential factors and significantly improves upon the nonadaptive GD rate in Theorem 1.

Theorem 2 (↓). *Let Assumption 1 hold and let the iterates $\{x_k\}_{k=0}^K$ be generated by eq. (8). Let the stepsizes $\eta_k > 0$ be defined according to the following options:*

$$\textit{Option A: } \eta_k = \mathcal{R}/[\|\nabla f(x_k)\|_2\sqrt{K+1}], \quad \textit{Option B: } \eta_k = [f(x_k) - f^*]/\|\nabla f(x_k)\|_2^2, \tag{11}$$

and let the number of iterations $K \in \mathbb{N}$ be defined as follows:

$$K = 4 \cdot \left\lceil [\mathcal{R}\mathcal{M}_1]^2 + [\mathcal{R}\mathcal{M}_0/\epsilon]^2 \right\rceil. \tag{12}$$

Then, the inequality $\min_{k\in\{0,\ldots,K\}}[f(x_k) - f^*] \leq \epsilon$ holds.

3.2 Stochastic Algorithms

In this section, we analyze the convergence of projected stochastic gradient descent (SGD):

$$x_{k+1} = \text{proj}_{Q_{\mathcal{R}}}(x_k - \eta_k \nabla_{\xi_k} f(x_k)), \quad x_0 = 0, \tag{13}$$

where $\xi_k \sim \mathcal{D}$ are independent samples and $\nabla_\xi f(x)$ is the stochastic gradient estimator defined in Assumption 2, a stochastic version of Assumption 1. The unbiasedness requirement is standard in stochastic first-order methods. The decomposition in eq. (14) can be viewed as a stochastic analogue of condition (1); in fact, it implies condition (1) with $\mathcal{M}_0 = \mathcal{G}_0$ and $\mathcal{M}_1 = \mathcal{G}_1$, as stated in Remark 1. For technical reasons, the condition in eq. (14) is slightly stronger than merely imposing an upper bound, analogous to condition (1), on the root second moment $\sqrt{\mathbb{E}[\|\nabla_\xi f(x)\|_2^2]}$. Nevertheless, it remains substantially more general than the standard bounded-second-moment assumption recovered by setting $\mathcal{G}_1 = 0$, which implies bounded full gradients and is commonly used in the stochastic non-smooth convex optimization literature (Nemirovski et al., 2009; Bubeck, 2015).

Assumption 2. *There exists a stochastic estimator $\nabla_\xi f(x)$ of the gradient $\nabla f(x)$, where $\xi \sim \mathcal{D}$ is a random variable sampled from the distribution $\mathcal{D}$. The gradient estimator $\nabla_\xi f(x)$ is unbiased, i.e., $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla_\xi f(x)] = \nabla f(x)$, and allows the decomposition $\nabla_\xi f(x) = u_\xi(x) + v_\xi(x)$, satisfying*

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|u_\xi(x)\|_2^2] \le \mathcal{G}_0^2 \quad \text{and} \quad \|v_\xi(x)\|_2 \overset{a.s.}{\le} \mathcal{G}_1[f(x) - f^*], \quad \text{where } \mathcal{G}_0, \mathcal{G}_1 \ge 0. \tag{14}$$

Remark 1. *Assumption 2 implies Assumption 1 with $\mathcal{M}_0 = \mathcal{G}_0$ and $\mathcal{M}_1 = \mathcal{G}_1$.*

Next, under Assumption 2, we analyze SGD with constant stepsizes and AdaGrad-Norm stepsizes (Streeter & McMahan, 2010), a scalar variant of AdaGrad. The corresponding results are stated in Theorems 3 and 4, respectively. The convergence rates for the two algorithms coincide. This is not particularly surprising: it is consistent with existing theory for stochastic convex globally Lipschitz optimization, where the rates for SGD and AdaGrad-Norm also coincide (Nemirovski et al., 2009; Duchi et al., 2011).

Theorem 3 (↓). *Let Assumption 2 hold and let the iterates $\{x_k\}_{k=0}^K$ be generated by eq. (13). Let the stepsizes $\eta_k > 0$ and the number of iterations $K \in \mathbb{N}$ be defined as follows:*

$$\eta_k = 1/\max\{2\mathcal{G}_1^2\mathcal{F}, \mathcal{G}_0\sqrt{K+1}/\mathcal{R}\}, \quad K = \mathcal{O}\big[\exp(\mathcal{O}[\mathcal{R}\mathcal{G}_1]) \cdot [\mathcal{R}\mathcal{G}_0/\epsilon] + [\mathcal{R}\mathcal{G}_0/\epsilon]^2\big]. \tag{15}$$

Then, the inequality $\mathbb{E}[f(\overline{x}_K) - f^*] \le \epsilon$ holds, where $\overline{x}_K = \frac{1}{K+1}\sum_{k=0}^K x_k$.

Theorem 4 (↓). *Let Assumption 2 hold and let the iterates $\{x_k\}_{k=0}^K$ be generated by eq. (13). Let the stepsizes $\eta_k > 0$ and the number of iterations $K \in \mathbb{N}$ be defined as follows:*

$$\eta_k = \mathcal{R}/\sqrt{2\textstyle\sum_{i=0}^k \|\nabla_{\xi_i} f(x_i)\|_2^2}, \quad K = \mathcal{O}\big[\exp(\mathcal{O}[\mathcal{R}\mathcal{G}_1]) \cdot [\mathcal{R}\mathcal{G}_0/\epsilon] + [\mathcal{R}\mathcal{G}_0/\epsilon]^2\big]. \tag{16}$$

Then, the inequality $\mathbb{E}[f(\overline{x}_K) - f^*] \le \epsilon$ holds, where $\overline{x}_K = \frac{1}{K+1}\sum_{k=0}^K x_k$.

As in the deterministic GD result in Theorem 1, the iteration complexities in Theorems 3 and 4 contain exponential factors and therefore can be very large. This raises the natural question of whether gradient normalization or Polyak stepsizes can yield improvements analogous to those obtained in the deterministic setting in Theorem 2. Unfortunately, the analysis in Theorem 2 does not directly extend to the stochastic setting: gradient normalization and Polyak stepsizes generally break the unbiasedness property, which is crucial for the convergence analysis. In fact, normalized SGD can diverge even on simple one-dimensional convex problems (Karimireddy et al., 2019, Counterexample 1). However, in the next Section 4, we show that it is still possible to significantly improve upon the convergence rates of SGD and AdaGrad-Norm.

4 Better Algorithms for Stochastic Convex Generalized Lipschitz Problems

4.1 Algorithm Construction

In this section, we show how to construct an algorithm with convergence guarantees that significantly improve upon those in Section 3.2. We begin by defining the functions $f_k(x)\colon \mathbb{R}^d \to \mathbb{R}$ for $k \in \mathbb{N}$:

$$f_k(x) = \pi_k f(\alpha_k x + (1-\alpha_k)x_{k-1}), \quad \text{where} \quad \pi_k(1-\alpha_k) = \pi_{k-1}. \tag{17}$$

Algorithm 1 Clipped AdamW (Theorem 5)

$m_0 = 0, v_0 = 0, x_1 = 0$
for $k = 1, \dots, K-1$ **do**
  $\hat{g}_k = \nabla_{\xi_k} f(x_k)$, where $\xi_k \sim \mathcal{D}$
  $m_k = m_{k-1} + \hat{g}_k$
  $v_k = v_{k-1} + \|\hat{g}_k\|_2^2$
  $x_{k+1} = \frac{k}{k+1} x_k - \frac{\mathcal{R}}{k+1} \operatorname{clip}_2(m_k/\sqrt{v_k})$

Algorithm 2 Clipped AdamW (Theorem 6)

$m_0 = 0, v_0 = 0, x_1 = 0$
for $k = 1, \dots, K-1$ **do**
  $\hat{g}_k = \nabla_{\xi_k} f(x_k)$, where $\xi_k \sim \mathcal{D}$
  $m_k = (1-\alpha) m_{k-1} + \alpha \hat{g}_k$
  $v_k = (1-\alpha)^2 v_{k-1} + \alpha^2 \|\hat{g}_k\|_2^2$
  $x_{k+1} = (1-\alpha) x_k - \alpha \mathcal{R} \operatorname{clip}_2(m_k/\sqrt{v_k})$

Here, $\alpha_k \in [0,1]$ and $\pi_{k-1} \geq 0$ for $k \in \mathbb{N}$. We then consider the following iterations for $k \in \mathbb{N}$:

$$x_k = (1-\alpha_k) x_{k-1} + \alpha_k z_k, \quad x_0 = 0, \tag{18}$$

where the points $z_k \in Q_{\mathcal{R}}$ are generated by an online optimization algorithm applied to an online linear optimization problem with regret

$$\mathrm{Reg}_K = \max_{z \in Q_{\mathcal{R}}} \sum_{k=1}^K \langle z_k - z, g_k \rangle, \quad \text{where} \quad g_k = \nabla_{\xi_k} f_k(z_k). \tag{19}$$

Here, $\xi_k \sim \mathcal{D}$ are random samples independent of the iterates $z_1, \dots, z_k$. When $1/\alpha_k = \pi_k = k$, the iterations in eq. (18) can be seen as a form of primal averaging (Nesterov & Shikhman, 2015; Tao et al., 2018) or online-to-batch conversion (Cutkosky, 2019). These techniques yield convergence guarantees for the last iterate x_K, rather than for the averaged iterate $\overline{x}_K$ as in Theorems 3 and 4. However, we will later show that this choice of parameters leads to significantly suboptimal convergence rates under Assumption 2. Consequently, in Lemma 2, we analyze the iterations in eq. (18) for a general choice of parameters.

Lemma 2 (↓). *Under Assumption 2, the following inequality holds:*

$$\pi_K \mathbb{E}[f(x_K) - f^*] \leq \pi_0 [f(x_0) - f^*] + \mathbb{E}[\mathrm{Reg}_K]. \tag{20}$$

We further focus on online optimization algorithms that guarantee the regret bound stated in Assumption 3. This bound is optimal when no assumptions are imposed on the gradients g_k (Orabona & Pál, 2018). Moreover, we will show that it enables adaptation to the unbounded gradients allowed under Assumption 2. Several algorithms achieve this bound, including AdaGrad-Norm (Streeter & McMahan, 2010), SOLO-FTRL (Orabona & Pál, 2018), and the scalar-stepsize variant of Leon (Jiang et al., 2026; Kovalev, 2026).

Assumption 3. *The iterates $z_k \in Q_{\mathcal{R}}$ satisfy the following, where $\mathcal{C} \geq 1$ is a universal constant:*

$$\mathrm{Reg}_K \overset{a.s.}{\leq} \mathcal{C} \mathcal{R} \sqrt{\textstyle\sum_{k=1}^K \|g_k\|_2^2}. \tag{21}$$

4.2 Convergence Analysis with Default Averaging

In this section, we analyze the convergence of the iterations in eq. (18) under the default choice of parameters: $1/\alpha_k = \pi_k = k$. When SOLO-FTRL is used as the online learning algorithm to guarantee the regret bound in Assumption 3, we obtain Algorithm 1, which is very similar to AdamW with scalar stepsizes and clipped updates. Further, we mostly focus on this particular algorithm, although the results and discussions apply to the iterations in eq. (18) combined with other online learning algorithms that generate the iterates z_k satisfying Assumption 3.

We begin the convergence analysis with an upper bound on the expected regret in Lemma 3. Next, in Theorem 5, we establish the main convergence result under the default choice of parameters. Finally, using the inequality in eq. (7), we derive an explicit iteration complexity in Corollary 2.

Lemma 3 (↓). *Let Assumptions 2 and 3 hold. Then, the following inequality holds:*

$$\mathbb{E}[\mathrm{Reg}_K] \leq \mathcal{C} \mathcal{R} \mathcal{G}_0 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathbb{E}\left[\mathcal{C} \mathcal{R} \mathcal{G}_1 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 [f(x_k) - f^*]^2} \right]. \tag{22}$$

Theorem 5 (↓). *Let Assumptions 2 and 3 hold, and assume $\mathcal{R}\mathcal{G}_1 \geq 1$. Let $\alpha_k = 1/k$ and $\pi_0 = 0$. Let $K \in \mathbb{N}$ be defined as follows:*

$$K = \left\lceil [\mathcal{R}\mathcal{G}_0/\epsilon]^2 + \left[[\mathcal{R}\mathcal{G}_1]^2 \cdot [\mathcal{F}/\epsilon] \right] \cdot \ln\left[e + [\mathcal{R}\mathcal{G}_0/\epsilon] + [\mathcal{R}\mathcal{G}_1]^2 \cdot [\mathcal{F}/\epsilon] \right] \right\rceil. \tag{23}$$

Then, the inequality $\mathbb{E}[f(x_K) - f^*] \leq \mathcal{O}[\epsilon]$ holds.

Corollary 2. *Under the conditions of Theorem 5, to reach the precision $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$, it is sufficient to perform the following number of iterations:*

$$K = \tilde{\mathcal{O}}\big[\exp(\mathcal{O}[\mathcal{R}\mathcal{G}_1]) \cdot [\mathcal{R}\mathcal{G}_0/\epsilon] + [\mathcal{R}\mathcal{G}_0/\epsilon]^2\big]. \tag{24}$$

Unfortunately, the complexity bound for Algorithm 1 in Corollary 2 does not improve upon the corresponding bounds for SGD and AdaGrad-Norm in Theorems 3 and 4. As in the discussion of GD convergence in Section 3.1, this bound admits an informal two-phase interpretation, with slow sublinear rates in both phases. The main limiting factor is the default choice of parameters for the iterations in eq. (18), which corresponds to simple averaging of the gradients and squared gradient norms in Lines 4 and 5 of Algorithm 1. This simple averaging is the key difference from practical Adam-type methods, which use exponential decay instead. In the next Section 4.3, we show that exponential decay is not merely a practical heuristic, but can substantially improve upon the complexity bound in Corollary 2.

4.3 Improved Rates via Exponential Decay

In this section, we show that replacing the simple averaging in Algorithm 1 with exponential decay improves the convergence rate. Specifically, we replace the default schedule $1/\alpha_k = \pi_k = k$ in eq. (18) with a constant parameter $\alpha_k = \alpha$. This yields Algorithm 2, which coincides with scalar-stepsize AdamW with clipped updates. The constant choice of α_k is exactly what induces the exponential decay in Lines 4 and 5 of Algorithm 2.

We begin the convergence analysis with Lemma 4, which gives the key improved upper bound on the regret. It exploits the fact that, with a sufficiently small choice of the parameters α_k, the iterates in eq. (18) remain stable, i.e., $x_k \approx x_{k-1}$. Intuitively, this may imply the stability of the gradients, i.e., $g_k \approx g_{k-1}$, which makes the method behave more like normalized GD in Theorem 2 by accumulating stable squared gradient norms in Line 5. Formally, the argument uses the stability of the function values $f(x_k)$, which accounts for the improvement of the bound in Lemma 4 over the bound in Lemma 3. Note that Nesterov & Shikhman (2015); Cutkosky (2019) also made informal stability-related observations about primal averaging and online-to-batch conversion, but did not provide any formal justification that these schemes actually improve convergence rates. In contrast, our analysis in Section G.5 is substantially different: it formally exploits the iterate stability and leads to improved convergence guarantees.

Lemma 4 (↓). *Let Assumptions 2 and 3 hold and assume $\mathcal{R}\mathcal{G}_1 \geq 1$. Let α_k satisfy the following:*

$$\alpha_k = \alpha, \quad \textit{where} \quad \alpha T \mathcal{R}\mathcal{G}_1 \leq 1 \quad \textit{and} \quad T = \big\lceil \hat{\mathcal{C}}[\mathcal{R}\mathcal{G}_1]^2 \big\rceil, \tag{25}$$

where $\hat{\mathcal{C}}$ is a sufficiently large universal constant. Then, the following inequality holds for all $K \geq T$:

$$\mathbb{E}[\mathrm{Reg}_K] \leq \tfrac{1}{2}\textstyle\sum_{k=1}^K \alpha_k \pi_k \mathbb{E}[f(x_k) - f^*] + \mathcal{O}\Big[\mathcal{R}\mathcal{G}_0 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + T\mathcal{R}\mathcal{G}_0 \textstyle\sum_{k=1}^K \alpha_k^2 \pi_k\Big]. \tag{26}$$

Armed with Lemmas 2 and 4, we derive the main convergence result in Theorem 6 and the explicit iteration complexity in Corollary 3. As for Algorithm 1 in Section 4.2, the complexity bound for Algorithm 2 in Corollary 3 admits a two-phase interpretation. However, while Algorithm 1 and the other stochastic methods from Section 3.2 achieve slow sublinear convergence rates in both phases, Algorithm 2 can achieve fast linear convergence in the first phase, up to the precision $\hat{\epsilon} = \mathcal{O}[\mathcal{G}_0/\mathcal{G}_1]$. This improvement can be significant when $\mathcal{R}\mathcal{G}_1 \gg 1$ and, more importantly, eliminates exponential factors from the iteration complexity.

Theorem 6 (↓). *Let Assumptions 2 and 3 hold, and assume $\mathcal{R}\mathcal{G}_1 \geq 1$. Let $\pi_0 = 1$, and let α_k and $K \in \mathbb{N}$ be defined as follows:*

$$1/\alpha_k = 1/\alpha = \max\big\{T\mathcal{R}\mathcal{G}_1, [\mathcal{R}\mathcal{G}_0/\epsilon]^2, [T\mathcal{R}\mathcal{G}_0/\epsilon]\big\}, \quad K = \big\lceil [2/\alpha_1] \ln[e + [\mathcal{F}/\epsilon]] \big\rceil, \tag{27}$$

where T is defined in eq. (25). Then, the inequality $\mathbb{E}[f(x_K) - f^*] \leq \mathcal{O}[\epsilon]$ holds.

Corollary 3. *Under the conditions of Theorem 6, to reach the precision $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$, it is sufficient to perform the following number of iterations:*

$$K = \tilde{\mathcal{O}}\big[[\mathcal{R}\mathcal{G}_1]^4 + [\mathcal{R}\mathcal{G}_1]^3 \cdot [\mathcal{R}\mathcal{G}_0/\epsilon] + \mathcal{R}\mathcal{G}_1 \cdot [\mathcal{R}\mathcal{G}_0/\epsilon]^2\big] \tag{28}$$

Unfortunately, a constant choice of α_k is no longer optimal when the target precision is below $\hat{\epsilon}$. However, this limitation can be addressed with a simple two-stage schedule for α_k. We extend the regret bound in eq. (26) to this schedule in Lemma 5, and then derive the corresponding convergence result and explicit complexity bound in Theorem 7 and Corollary 4. The resulting guarantee combines the best features of Algorithms 1 and 2: it achieves fast linear convergence in the first phase and optimal sublinear convergence in the second phase.

Lemma 5 (↓). *Under the conditions of Theorem 7, the inequality in eq. (26) holds for $K \geq T$.*

Theorem 7 (↓). *Let Assumptions 2 and 3 hold, and assume $\mathcal{RG}_1 \geq 1$. Let α_k be defined as follows:*

$$1/\alpha_k = \begin{cases} T\mathcal{RG}_1 & k \leq ST \\ \max\{T\mathcal{RG}_1, [\mathcal{RG}_0/\epsilon]^2, [T\mathcal{RG}_0/\epsilon]\} & k > ST \end{cases}, \tag{29}$$

where T is defined in eq. (25), and let $\pi_0 = 1$. Let $K, S \in \mathbb{N}$ be defined as follows:

$$K = TS + \lceil [2/\alpha_{TS+1}] \ln[e + [\mathcal{G}_0/(\mathcal{G}_1\epsilon)]] \rceil, \quad S = \lceil [2/(\alpha_1 T)] \ln[e + [\mathcal{FG}_1/\mathcal{G}_0]] \rceil. \tag{30}$$

Then, the inequality $\mathbb{E}[f(x_K) - f^*] \leq \mathcal{O}[\epsilon]$ holds.

Corollary 4. *Under the conditions of Theorem 7, to reach the precision $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$, it is sufficient to perform the following number of iterations:*

$$K = \tilde{\mathcal{O}}\big[[\mathcal{RG}_1]^4 + [\mathcal{RG}_0/\epsilon]^2\big]. \tag{31}$$

4.4 Discussion

We conclude this section by summarizing the results obtained above and discussing them in the context of modern literature.

Advanced vs classical averaging. Several advanced averaging schemes for convex optimization have been developed in recent years, including primal averaging (Nesterov & Shikhman, 2015; Tao et al., 2018), online-to-batch conversion (Cutkosky, 2019), and other schemes (Defazio et al., 2024). Although these schemes are aesthetically appealing (Cutkosky, 2019) and can improve practical performance (Defazio et al., 2024), there is currently no theoretical evidence that they can lead to improved convergence rates compared to the classical Polyak–Ruppert averaging (Polyak, 1990; Ruppert, 1988) which is used, for instance, in Theorems 3 and 4; refer to Defazio et al. (2024, eqs. 1 and 2) for the definition. To the best of our knowledge, Theorem 7 is the first result of this kind: it shows that the averaging scheme in eq. (18) yields significantly improved convergence guarantees under Assumption 2, which allows unbounded gradients.

Exponential decay in Adam/AdamW. The original Adam paper (Kingma & Ba, 2014) interprets the exponentially weighted accumulations in Lines 4 and 5 of Algorithm 2 as estimates of the first and second moments of the gradient obtained via exponential moving averages. Unfortunately, this interpretation does not provide meaningful insight into the convergence properties of Adam/AdamW.

An alternative interpretation was proposed by Défossez et al. (2020) and later revisited by Chezhegov et al. (2024). To analyze the convergence of Adam, they compare AdaGrad with its counterpart, RMSProp (Tieleman, 2012), which replaces AdaGrad's cumulative sum of squared gradients (see, e.g., eq. (16)) with an exponentially weighted sum. Défossez et al. (2020, Section 4.3) suggest that RMSProp relates to AdaGrad in the same way that constant-stepsize SGD relates to decaying-stepsize SGD. This interpretation yields meaningful convergence guarantees for RMSProp; however, these guarantees do not improve upon those for AdaGrad and therefore do not explain the benefits of exponential decay. In contrast, our analysis in Section 4.3 shows that understanding Adam or AdamW through the RMSProp–AdaGrad analogy may be inaccurate: in the unbounded-gradient setting covered by Assumption 2, AdamW obtains the improved convergence guarantees in Theorem 7 by essentially exploiting iterate stability, which is absent in AdaGrad and RMSProp.

Non-convex analysis of Adam. Another, more accurate interpretation of Adam with exponential decay was proposed by Ahn et al. (2024); Ahn & Cutkosky (2024). They analyze the convergence of Adam for non-convex functions using the exponentiated online-to-non-convex conversion (Zhang & Cutkosky, 2024). Their approach shares some similarities with ours, as we also use a form of conversion from convex optimization to online optimization, as described in Section 4.1. However, their analysis does not show that Adam can improve upon the convergence rates of SGD with

momentum (Cutkosky et al., 2023). One reason is that SGD with momentum is already optimal for globally Lipschitz functions (Cutkosky et al., 2023), which highlights the importance of realistic assumptions like Assumption 2. Another reason is that their analysis is limited to general non-convex functions. Therefore, they can only obtain vague first-order stationarity guarantees, which typically depend on the initial function gap $f(x_0) - f^*$. This gap can be exponential due to Lemma 1, making exponential factors in the resulting iteration complexities unavoidable. In contrast, in Theorems 6 and 7, we obtain global convergence rates for convex functions with only logarithmic dependence on the function gap $\mathcal{F}$, which is essential for avoiding exponential factors in the complexity and improving upon other methods. We highlight that such theoretical results for convex functions are relevant to machine learning practice, as convexity-like properties have been observed in neural-network training despite global nonconvexity (Kleinberg et al., 2018; Zhou et al., 2019; Tran et al., 2024). We also extend our results to the quasar-convex setting (Hinder et al., 2020) in Section C, bringing them even closer to practice.

It is also worth noting that general non-convex analysis can limit the possibility of understanding why Adam outperforms other methods even without assuming that the objective function is globally Lipschitz. For instance, Li et al. (2023b) derived the convergence guarantees for Adam on non-convex functions under the generalized smoothness assumption, but acknowledged that their results do not improve upon the convergence rate of SGD (Li et al., 2023a). In contrast, in Section B, we derive convergence guarantees for Algorithm 2 on convex functions under generalized Hölder smoothness, and these guarantees also improve upon the results for other methods.

Non-monotone stepsizes. It is well known that AdaGrad has the limitation that its stepsizes are always non-increasing (Defazio et al., 2022). On the other hand, the stepsize in Line 6 of Algorithm 2 is proportional to $1/\sqrt{v_k}$ and can be non-monotone due to the exponential decay in Line 5. Although common intuition suggests that such non-monotonicity can be beneficial because it allows the method to "forget" gradients that are too "old", to the best of our knowledge, there are no theoretical results supporting this intuition. Our results in Theorems 6 and 7 support this intuition: the stepsize in Algorithm 2 may be exponentially small in early iterations, because the gradients may be exponentially large due to the bound in eq. (7); however, exponential decay allows the method to adapt to these large gradients and yields convergence rates without exponential factors, as discussed in Section 4.3.

Reliance on the new assumption. Our theoretical analysis is developed under the newly introduced generalized Lipschitzness assumption. However, the conclusions of this section are not specific to this function class and also hold under generalized smoothness, a well-established and extensively studied assumption (Zhang et al., 2019). In particular, we extend the convergence analysis from Section 4 to this setting in Section B. The remaining results in Table 1 and the lower bound constructions in Section A extend straightforwardly; some of these extensions are already available in the literature (Li et al., 2023a; Gorbunov et al., 2024; Vankov et al., 2024). We deliberately leave out these extensions to avoid unnecessary technical overcomplication of our paper.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including grammar correction, removal of repetition, and rewording to improve clarity and readability. AI-assisted editing was limited to text segments no longer than a single paragraph at a time. All suggested edits were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. *Advances in Neural Information Processing Systems*, 37:94909–94933, 2024.

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise. *arXiv preprint arXiv:2402.01567*, 2024.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

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

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*, 2024.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

Ashok Cutkosky. Anytime online-to-batch, optimism and acceleration. In *International conference on machine learning*, pp. 1446–1454. PMLR, 2019.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670. PMLR, 2023.

Aaron Defazio, Baoyu Zhou, and Lin Xiao. Grad-gradagrad? a non-monotone adaptive stochastic gradient method. *arXiv preprint arXiv:2206.06900*, 2022.

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

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

John W Hagood and Brian S Thomson. Recovering a function from a dini derivative. *The American Mathematical Monthly*, 113(1):34–46, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pp. 1894–1938. PMLR, 2020.

Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. *arXiv preprint arXiv:2602.08232*, 2026.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International conference on machine learning*, pp. 3252–3261. PMLR, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pp. 2698–2707. PMLR, 2018.

Dmitry Kovalev. Sgd with adaptive preconditioning: Unified analysis and momentum acceleration. *arXiv preprint arXiv:2506.23803*, 2025.

Dmitry Kovalev. Optimal projection-free adaptive sgd for matrix optimization. *arXiv preprint arXiv:2604.02505*, 2026.

Dmitry Kovalev and Ekaterina Borodich. Non-euclidean sgd for structured optimization: Unified analysis and improved rates. *arXiv preprint arXiv:2511.11466*, 2025.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023a.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023b.

Aleksandr Lobanov and Anastasia Koloskova. Avoiding bias in clipped sgd for overparameterized models under generalized smoothness. *arXiv preprint arXiv:2605.14800*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

Yu Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.

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

Francesco Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

Boris T Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

Boris Teodorovich Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

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

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

Anton Rodomanov, Xiaowen Jiang, and Sebastian Stich. Universality of adagrad stepsizes for stochastic optimization: Inexact oracle, acceleration and variance reduction. *Advances in Neural Information Processing Systems*, 37:26770–26813, 2024a.

Anton Rodomanov, Ali Kavis, Yongtao Wu, Kimon Antonakopoulos, and Volkan Cevher. Universal gradient methods for stochastic convex optimization. *arXiv preprint arXiv:2402.03210*, 2024b.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Alexander Shapiro. On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487, 1990.

Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*. Springer Science & Business Media, 2012.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Wei Tao, Zhisong Pan, Gaowei Wu, and Qing Tao. Primal averaging: A new gradient evaluation step to attain the optimal individual convergence. *IEEE transactions on cybernetics*, 50(2):835–845, 2018.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

Hoang Tran, Qinzi Zhang, and Ashok Cutkosky. Reevaluating theoretical analysis methods for optimization in deep learning. *arXiv preprint arXiv:2407.01825*, 2024.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800*, 2024.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. *arXiv preprint arXiv:2405.09742*, 2024.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

A LOWER BOUNDS

The complexity results for SGD and AdaGrad-Norm established in Section 3.2 for stochastic convex optimization under Assumption 2 are quite pessimistic. This naturally raises the question of whether the exponential factors in the complexities obtained in Theorems 3 and 4 are unavoidable or merely artifacts of a loose convergence analysis. In Theorems 8 and 9, we provide lower bounds showing that these factors are indeed unavoidable. Therefore, in view of the improved complexity bound in Corollary 4, we conclude that scalar-stepsize AdamW with clipped updates provably outperforms SGD and AdaGrad in this setting.

Theorem 8 (↓). *Let the constants $\mathcal{G}_0, \mathcal{G}_1, \mathcal{R}, \epsilon > 0$ satisfy the following conditions:*

$$\epsilon \leq [\mathcal{G}_0/\mathcal{G}_1] \exp(\tfrac{1}{8}\mathcal{R}\mathcal{G}_1), \qquad \mathcal{R}\mathcal{G}_1 \geq 8. \tag{32}$$

Let $\eta_k = \eta > 0$ be an arbitrary stepsize, and let $\overline{x}_K = \frac{1}{K+1}\sum_{k=0}^{K} x_k$, where $x_0, \ldots, x_k$ are defined by eq. (13). There exists a convex function $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ and corresponding stochastic gradient oracle $\nabla_\xi f(x)\colon \mathbb{R}^d \to \mathbb{R}^d$ satisfying Assumption 2, such that at least the following number of iterations is required to reach the precision $\mathbb{E}[\min\{f(x_K), f(\overline{x}_K)\} - f^*] \leq \epsilon$:

$$K = \Omega[\exp(\Omega[\mathcal{R}\mathcal{G}_1]) \cdot [\mathcal{R}\mathcal{G}_0/\epsilon]]. \tag{33}$$

Theorem 9 (↓). *Let the constants $\mathcal{G}_0, \mathcal{G}_1, \mathcal{R}, \epsilon > 0$ satisfy the following conditions:*

$$\epsilon \leq [\mathcal{G}_0/\mathcal{G}_1] \exp(\tfrac{1}{64}\mathcal{R}\mathcal{G}_1), \qquad \mathcal{R}\mathcal{G}_1 \geq 32. \tag{34}$$

Let $\eta > 0$ be an arbitrary parameter, and let $\overline{x}_K = \frac{1}{K+1}\sum_{k=0}^{K} x_k$, where $x_0, \ldots, x_k$ are defined by eq. (13) with the AdaGrad-Norm stepsizes:

$$\eta_k = \eta/\sqrt{\textstyle\sum_{i=0}^{k} \|\nabla_{\xi_i} f(x_i)\|_2^2} \tag{35}$$

There exists a convex function $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ and corresponding stochastic gradient oracle $\nabla_\xi f(x)\colon \mathbb{R}^d \to \mathbb{R}^d$ satisfying Assumption 2, such that at least the following number of iterations is required to reach the precision $\mathbb{E}[\min\{f(x_K), f(\overline{x}_K)\} - f^*] \leq \epsilon$:

$$K = \Omega[\exp(\Omega[\mathcal{R}\mathcal{G}_1])]. \tag{36}$$

B UNIVERSALITY: GENERALIZED HÖLDER SMOOTH STOCHASTIC CONVEX PROBLEMS

B.1 GENERALIZED HÖLDER SMOOTHNESS

In this section, we assume that the objective function $f(x)$ is differentiable and convex. Next, we introduce Assumption 4, which we call stochastic generalized Hölder smoothness.

Assumption 4. *There exists a stochastic estimator $\nabla_\xi f(x)$ of the gradient $\nabla f(x)$, where $\xi \sim \mathcal{D}$ is a random variable sampled from the distribution $\mathcal{D}$. The gradient estimator $\nabla_\xi f(x)$ is unbiased, i.e., $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla_\xi f(x)] = \nabla f(x)$, and allows the decomposition $\nabla_\xi f(x) = u_\xi(x) + v_\xi(x)$, satisfying*

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|u_\xi(x)\|_2^2] \le \sigma^2 + \mathcal{L}_0^{\frac{2}{1+\nu}}[f(x) - f^*]^{\frac{2\nu}{1+\nu}} \quad \text{and} \quad \|v_\xi(x)\|_2 \overset{a.s.}{\le} \mathcal{L}_1[f(x) - f^*], \tag{37}$$

where $\nu \in [0,1]$ and $\sigma, \mathcal{L}_0, \mathcal{L}_1 \ge 0$, and where $\sigma = 0$ whenever $\nu = 0$.

We motivate Assumption 4 by describing the following special cases.

- **(i)** In the case $(\nu, \sigma, \mathcal{L}_0, \mathcal{L}_1) = (0, 0, \mathcal{G}_0, \mathcal{G}_1)$, Assumption 4 matches Assumption 2, which is a stochastic variant of the generalized Lipschitzness discussed in Section 2.
- **(ii)** In the noiseless case with $(\nu, \sigma) = (1, 0)$, Assumption 4 recovers the generalization of the $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness up to universal constants; see eq. (4) in Section 2.
- **(iii)** In the noiseless case with $(\sigma, \mathcal{L}_0, \mathcal{L}_1) = (0, \mathcal{L}, 0)$, Assumption 4 is implied by the standard $(\nu, \mathcal{L})$-Hölder smoothness assumption (Nesterov, 2015) up to universal constants; see Corollary 1 of Orabona (2023).
- **(iv)** Assuming the stochastic setting with globally bounded variance of the stochastic gradient and allowing $\sigma \neq 0$ in cases **(ii)** and **(iii)** recovers the stochastic variants of the corresponding assumptions.

By analyzing these cases, one can conclude that Assumption 4 interpolates between the stochastic variants of the generalized Lipschitzness and generalized smoothness, just like the standard Hölder smoothness interpolates between uniform Lipschitzness and uniform smoothness.

We also establish an important property of the generalized Hölder smooth functions in the following Lemma 6. It can be seen as the extension of Lemma 1 to the stochastic generalized Hölder smoothness setting.

Lemma 6 (↓). *Let Assumption 4 hold and let the function $h(x)\colon \mathbb{R}^d \to \mathbb{R}$ be defined as follows:*

$$h(x) = [\beta + f(x) - f^*]^{\frac{1}{1+\nu}}, \quad \text{where} \quad \beta = \min\left\{[\sigma^{1+\nu}/\mathcal{L}_0]^{\frac{1}{\nu}}, \sigma/\mathcal{L}_1\right\} \tag{38}$$

Then, the following inequality holds for all $x, x' \in \mathbb{R}^d$:

$$|h(x') - h(x)|^{1+\nu} \le \mathcal{O}\left[\left(\mathcal{L}_0 + [\mathcal{L}_1 h(x)]^{1+\nu}\right)\exp(\mathcal{L}_1\|x' - x\|_2)\|x' - x\|_2^{1+\nu}\right]. \tag{39}$$

B.2 CONVERGENCE ANALYSIS

In this section, we analyze the convergence of the iterates generated by eq. (18) under Assumptions 3 and 4. Our goal is to extend the two-stage convergence guarantee of Theorem 7 to the stochastic generalized Hölder smooth setting. We first prove the corresponding generalized regret bound in Lemma 7, which extends eq. (26), and then derive the main convergence result and explicit iteration complexity bound in Theorem 10 and Corollary 5.

Lemma 7 (↓). *Under the conditions of Theorem 10, the following inequality holds for all $K \ge T$:*

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_K] \le{}& \tfrac{1}{2}\textstyle\sum_{k=1}^K \alpha_k\pi_k\mathbb{E}[f(x_k) - f^*] + \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + T\mathcal{R}\sigma\textstyle\sum_{k=1}^K \alpha_k^2\pi_k\right] \\ &+ \mathcal{O}\left[\mathcal{R}^{1+\nu}\mathcal{L}_0\left[\textstyle\sum_{k=1}^K [\alpha_k\pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\textstyle\sum_{k=1}^K \alpha_k^{2+\nu}\pi_k\right].\end{aligned} \tag{40}$$

Theorem 10 (↓). *Let Assumptions 3 and 4 hold, and assume $\mathcal{R}\mathcal{L}_1 \ge 1$. Let $T = \left\lceil \hat{\mathcal{C}}[\mathcal{R}\mathcal{L}_1]^2\right\rceil$, where $\hat{\mathcal{C}}$ is a sufficiently large universal constant, and let $\pi_0 = 1$. Let α_k be defined as follows:*

$$\frac{1}{\alpha_k} = \begin{cases} T\mathcal{R}\mathcal{L}_1 & k \le ST \\ \max\left\{T\mathcal{R}\mathcal{L}_1, [\mathcal{R}\sigma/\epsilon]^2, \left[\mathcal{R}^{1+\nu}\mathcal{L}_0/\epsilon\right]^{\frac{2}{1+\nu}}, [T\mathcal{R}\sigma/\epsilon], \left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0/\epsilon\right]^{\frac{1}{1+\nu}}\right\} & k > ST \end{cases}. \tag{41}$$

Let $\hat{\mathcal{F}} = \mathcal{L}_0/\mathcal{L}_1^{1+\nu} + \sigma/\mathcal{L}_1$*, and let* $K, S \in \mathbb{N}$ *be defined as follows:*

$$K = TS + \lceil [2/\alpha_{TS+1}] \ln[e + [\hat{\mathcal{F}}/\epsilon]] \rceil, \quad S = \lceil [2/(\alpha_1 T)] \ln[e + [\mathcal{F}/\hat{\mathcal{F}}]] \rceil. \tag{42}$$

Then, the inequality $\mathbb{E}[f(x_K) - f^*] \leq \mathcal{O}[\epsilon]$ *holds.*

Corollary 5. *Under the conditions of Theorem 10, to reach the precision* $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$*, it is sufficient to perform the following number of iterations:*

$$K = \tilde{\mathcal{O}}\big([\mathcal{R}\mathcal{L}_1]^4 + [\mathcal{R}^{1+\nu}\mathcal{L}_0/\epsilon]^{\frac{2}{1+\nu}} + [\mathcal{R}\sigma/\epsilon]^2\big]. \tag{43}$$

As with the corresponding result for the generalized Lipschitz setting in Section 4.3, the guarantee in Corollary 5 admits an informal two-phase interpretation: fast linear convergence in the first phase up to the precision $\hat{\epsilon} = \mathcal{O}[\max\{\mathcal{R}\sigma/[\mathcal{R}\mathcal{L}_1]^2, \mathcal{R}^{1+\nu}\mathcal{L}_0/[\mathcal{R}\mathcal{L}_1]^{2(1+\nu)}\}]$, followed by the phase with the sublinear convergence rate, which matches the standard convergence rate of SGD for convex $(\nu, \mathcal{L}_0)$-Hölder smooth functions with bounded stochastic gradient variance (Rodomanov et al., 2024b). It is worth highlighting that concurrent work by Lobanov & Koloskova (2026) obtains a similar first-phase linear convergence result in the generalized smooth case $\nu = 1$; however, their guarantees rely on significantly stronger or even unrealistic assumptions, such as increasing batch sizes or the interpolation condition.

The result in Theorem 10 also implies that, up to logarithmic factors, the schedule of α_k can be parameterized by the overall number of iterations K and the constants $\mathcal{R}$ and $\mathcal{L}_1$, without prior knowledge of the Hölder exponent ν. This parameter independence across smoothness levels is called universality (Nesterov, 2015) and is also a feature of many other adaptive algorithms such as AdaGrad (Rodomanov et al., 2024a).

C Extension to Quasar-Convex Setting

C.1 Quasar Convexity and Generalized Gradients

In this section, we no longer assume the convexity of the objective function $f(x)$. Instead, we assume that the function $f(x)$ is γ-quasar convex for $\gamma \in (0, 1]$, that is, the following inequality holds:

$$f(tx^* + (1-t)x) \leq \gamma t f^* + (1 - \gamma t) f(x) \quad \text{for all} \quad x \in \mathbb{R}^d, \ t \in [0, 1]. \tag{44}$$

Refer to Hinder et al. (2020) for more details on quasar-convex functions. Consequently, the function $f(x)$ may no longer be subdifferentiable at every point $x \in \mathbb{R}^d$. However, we can define the generalized subdifferential $\partial^\circ f(x) = \partial \phi_x(0)$, where $\phi_x(w) = f^\circ(x; w)$. Here, $f^\circ(x; w)$ is the generalized directional derivative (Clarke, 1990), which is a convex function with respect to the second argument and is defined for all $x, w \in \mathbb{R}^d$ as follows:

$$f^\circ(x; w) = \limsup_{x' \to x, t \to +0} \tfrac{1}{t}[f(x' + tw) - f(x')]. \tag{45}$$

Note that both $\partial^\circ f(x)$ and $f^\circ(x; w)$ exist because the function $f(x)$ is assumed to be locally Lipschitz. Furthermore, we assume that the objective function $f(x)$ is directionally differentiable in the sense of Hadamard (Shapiro, 1990), that is, the following limit exists for all $x, w \in \mathbb{R}^d$:

$$f'(x; w) = \lim_{w' \to w, t \to +0} \tfrac{1}{t}[f(x + tw') - f(x)]. \tag{46}$$

Now, we are ready to describe the generalized gradient oracle that we will further use in this section. Specifically, we assume that there exists $\nabla f(x; w) \in \mathbb{R}^d$, which we refer to as the generalized gradient and which satisfies the following properties for all $x, w \in \mathbb{R}^d$:

$$\nabla f(x; w) \in \partial^\circ f(x) \quad \text{and} \quad \langle w, \nabla f(x; w) \rangle = f'(x; w). \tag{47}$$

The existence of such a generalized gradient is guaranteed by Zhang et al. (2020, Lemma 4). Refer to Zhang et al. (2020, Section 2) for more details on generalized gradients. Moreover, in the following Lemma 8, we show that the generalized gradient can be used to define quasar-convexity via the first-order criterion, just as it is for differentiable functions.

Lemma 8 (↓). *The inequality in eq. (44) is equivalent to the following condition:*

$$\gamma[f(x) - f^*] \leq \langle x - x^*, \nabla f(x; w) \rangle \quad \textit{for all} \quad x, w \in \mathbb{R}^d. \tag{48}$$

We also introduce the following Assumption 5, which is an extension of Assumption 2 for the generalized gradient setting.

Assumption 5. *There exists a stochastic estimator $\nabla_\xi f(x; w)$ of the generalized gradient $\nabla f(x; w)$, where $\xi \sim \mathcal{D}$ is a random variable sampled from the distribution $\mathcal{D}$. The generalized gradient estimator $\nabla_\xi f(x; w)$ is unbiased, i.e., $\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla_\xi f(x; w)] = \nabla f(x; w)$, and allows the decomposition $\nabla_\xi f(x; w) = u_\xi(x; w) + v_\xi(x; w)$, satisfying*

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\|u_\xi(x; w)\|_2^2] \leq \mathcal{G}_0^2 \quad \textit{and} \quad \|v_\xi(x; w)\|_2 \overset{a.s.}{\leq} \mathcal{G}_1[f(x) - f^*], \quad \textit{where} \ \ \mathcal{G}_0, \mathcal{G}_1 > 0. \tag{49}$$

Lemma 9. *Assumption 5 implies Assumption 1 with $\mathcal{M}_0 = \mathcal{G}_0$ and $\mathcal{M}_1 = \mathcal{G}_1$.*

C.2 Algorithm and Convergence Analysis

Similar to Section 4, we consider the iterations in eq. (18), where the iterates $z_k \in Q_{\mathcal{R}}$ are obtained by an online learning algorithm for bounding the regret Reg_K in eq. (19), satisfying Assumption 3, but with a different definition of g_k:

$$g_k = \alpha_k \pi_k \nabla_{\xi_k} f(\overline{x}_k; x_k - x_{k-1}), \quad \overline{x}_k = x_{k-1} + \zeta_k (x_k - x_{k-1}), \tag{50}$$

where the samples $\xi_k \sim \mathcal{D}$ and $\zeta_k \sim$ u.a.r.$([0, 1])$ are independent of each other and of $z_1, \ldots, z_k$. Here, the generalized stochastic gradient is computed at the point $\overline{x}_k$ sampled uniformly at random on the segment $[x_k, x_{k-1}]$ which is inspired by the approach of Zhang et al. (2020); Cutkosky et al. (2023) for non-smooth non-convex optimization. Below, we derive the main convergence result and explicit iteration complexity bound in Theorem 11 and Corollary 6.

Theorem 11 (↓). *Let Assumptions 3 and 5 hold and let $\mathcal{RG}_1 \geq 1$. Let π_k be defined as follows:*

$$\pi_0 = 1, \qquad (1 - \gamma\alpha_k)\pi_k = \pi_{k-1}. \tag{51}$$

Let $T = \lceil \hat{\mathcal{C}}[\mathcal{RG}_1/\gamma]^2 \rceil$ for a sufficiently large universal constant $\hat{\mathcal{C}}$, and let α_k be defined as follows:

$$1/\alpha_k = \begin{cases} T\mathcal{RG}_1 & k \leq ST \\ \max\{T\mathcal{RG}_1, [1/\gamma] \cdot [\mathcal{RG}_0/\epsilon]^2, [1/\gamma] \cdot [T\mathcal{RG}_0/\epsilon]\} & k > ST \end{cases}. \tag{52}$$

Let $K, S \in \mathbb{N}$ be defined as follows:

$$K = TS + \lceil [2/(\gamma\alpha_{TS+1})] \ln[e + [\mathcal{G}_0/(\gamma\mathcal{G}_1\epsilon)]] \rceil, \quad S = \lceil [2/(\gamma\alpha_1 T)] \ln[e + [\gamma\mathcal{F}\mathcal{G}_1/\mathcal{G}_0]] \rceil. \tag{53}$$

Then, the inequality $\mathbb{E}[f(x_K) - f^*] \leq \mathcal{O}[\epsilon]$ holds.

Corollary 6. *Under the conditions of Theorem 11, to reach the precision $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$, it is sufficient to perform the following number of iterations:*

$$K = \tilde{\mathcal{O}}\big[[1/\gamma]^6 \cdot [\mathcal{RG}_1]^4 + [1/\gamma]^2 \cdot [\mathcal{RG}_0/\epsilon]^2\big]. \tag{54}$$

Algorithm 3 AdamW/LeonW (Diagonal Preconditioning)

1: $m_0 = 0$, $v_0 = \delta \mathbf{1}_d$, $x_1 = 0$
2: **for** $k = 1, \ldots, K-1$ **do**
3: $\quad \hat{g}_k = \nabla_{\xi_k} f(x_k)$, where $\xi_k \sim \mathcal{D}$
4: $\quad m_k = (1-\alpha_k) m_{k-1} + \alpha_k \hat{g}_k, \quad v_k = (1-\alpha_k)^2 v_{k-1} + \alpha_k^2 (\hat{g}_k \odot \hat{g}_k)$
5: $\quad$ Option I (AdamW): $x_{k+1} = (1-\alpha_{k+1}) x_k - \alpha_{k+1} \mathcal{R}\, \mathrm{clip}_\infty(m_k \oslash \sqrt{v_k})$
6: $\quad$ Option II (LeonW): $x_{k+1} = (1-\alpha_{k+1}) x_k - \alpha_{k+1} \mathcal{R}(m_k \oslash \sqrt{m_k \odot m_k + v_k})$

Algorithm 4 LeonW (Matrix Preconditioning)

1: $M_0 = 0$, $V_0 = \delta I_{m \times m}$, $X_1 = 0$
2: **for** $k = 1, \ldots, K-1$ **do**
3: $\quad \hat{G}_k = \nabla_{\xi_k} f(X_k)$, where $\xi_k \sim \mathcal{D}$
4: $\quad M_k = (1-\alpha_k) M_{k-1} + \alpha_k \hat{G}_k, \quad V_k = (1-\alpha_k)^2 V_{k-1} + \alpha_k^2 \hat{G}_k \hat{G}_k^\top$
5: $\quad X_{k+1} = (1-\alpha_{k+1}) X_k - \alpha_{k+1} \mathcal{R}[M_k M_k^\top + V_k]^{\dagger/2} M_k$

D Beyond Scalar stepsizes

In this section, we discuss that the results obtained in the main part of the paper are not specific to scalar stepsizes. First, we consider Algorithm 3 with diagonal preconditioning, which has two options. Option I in Line 5 solves the online linear optimization problem in eq. (19)[1] using the diagonal version of SOLO-FTRL, which leads to AdamW with the standard diagonal preconditioning and clipped updates. Option II in Line 6 solves the problem in eq. (19)[1] using a diagonal variant of Leon and leads to a slight variation of AdamW without clipping, which we call LeonW. We also consider Algorithm 4 – a variant of LeonW with matrix preconditioning for solving an optimization problem over the space of matrices $\mathbb{R}^{m \times n}$. This algorithm solves the auxiliary online optimization problem using the matrix variant of Leon.

The convergence analysis of these algorithms can exploit possible structural properties of the objective function, such as different coordinate scales (Algorithm 3) or low-rank gradients (Algorithm 4). In fact, we can analyze a unification of Algorithms 3 and 4 with general preconditioners, based on the theoretical framework of Kovalev & Borodich (2025); Kovalev (2025; 2026). More concretely, the online optimization problem in eq. (19)[1] can be solved using algorithms with general preconditioning described by Kovalev (2025; 2026). The resulting convergence analysis will be almost identical to the analysis from this paper, with the main difference being the replacement of the standard Euclidean norm with the weighted Euclidean or non-Euclidean norms. Therefore, we deliberately leave out all the formal details about this extension to avoid unnecessary technical overcomplication of this paper.

[1] Here we should change the definition of the ball $Q_{\mathcal{R}}$ using an appropriate, possibly non-Euclidean, norm, e.g., infinity norm for diagonal preconditioning and spectral norm for matrix preconditioning.

E PROOFS FOR SECTION 2

E.1 PROOF OF LEMMA 1

Equation (5) implies Assumption 1. We can upper $\limsup_{u\to 0} \frac{1}{\|u\|_2}|f(x+u)-f(x)|$ as follows:

$$\begin{aligned}\limsup_{u\to 0} \frac{1}{\|u\|_2}|f(x+u)-f(x)| &\overset{(a)}{\leq} \limsup_{u\to 0}(\mathcal{M}_0+\mathcal{M}_1[f(x)-f^*])\exp(\mathcal{M}_1\|u\|_2)\\ &= \mathcal{M}_0+\mathcal{M}_1[f(x)-f^*],\end{aligned}$$

where (a) uses eq. (5).

Assumption 1 implies Equation (5). Let the function $p(t)\colon \mathbb{R}_+ \to \mathbb{R}$ be defined as follows:

$$p(t) = f(x+t(x'-x))-f^*,$$

let the function $q(t)\colon \mathbb{R}_+ \to \mathbb{R}$ be defined as follows:

$$q(t) = \|x'-x\|_2 \int_0^t (\mathcal{M}_0+\mathcal{M}_1 p(\tau))\mathrm{d}\tau,$$

and let the function $r(t)\colon \mathbb{R}_+ \to \mathbb{R}$ be defined as follows:

$$r(t) = |p(t)-p(0)| - q(t).$$

First, we need to show that $r(t)\leq 0$ for all $t\in\mathbb{R}_+$. Observe that $r(0)=0$. Hence, it sufficient to show that the function $r(t)$ is non-increasing. Let $r^+(t)$ be the upper Dini derivative of the function $r(t)$, which is defined as follows:

$$r^+(t) = \limsup_{t'\downarrow t} \frac{1}{t'-t}[r(t')-r(t)].$$

We can upper-bound $r^+(t)$ as follows:

$$\begin{aligned}r^+(t) &\overset{(a)}{=} \limsup_{t'\downarrow t} \frac{1}{t'-t}[|p(t')-p(0)|-q(t')-|p(t)-p(0)|+q(t)]\\ &\leq \limsup_{t'\downarrow t} \frac{1}{t'-t}[|p(t')-p(0)|-|p(t)-p(0)|] - \lim_{t'\to t}\frac{1}{t'-t}[q(t')-q(t)]\\ &\leq \limsup_{t'\downarrow t} \frac{1}{t'-t}|p(t')-p(t)| - \lim_{t'\to t}\frac{1}{t'-t}[q(t')-q(t)]\\ &\overset{(b)}{=} \limsup_{t'\downarrow t} \frac{1}{t'-t}|p(t')-p(t)| - \|x'-x\|_2(\mathcal{M}_0+\mathcal{M}_1 p(t))\\ &\leq \limsup_{t'\to t} \frac{1}{|t'-t|}|p(t')-p(t)| - \|x'-x\|_2(\mathcal{M}_0+\mathcal{M}_1 p(t))\\ &\overset{(c)}{\leq} 0,\end{aligned}$$

where (a) uses the definition of the functions $r(t)$ and $p(t)$; (b) uses the definition of the function $q(t)$ and the fact that the function $p(t)$ is continuous; (c) uses the definition of the function $p(t)$ and Assumption 1. Hence, by the Monotonicity Theorem (Hagood & Thomson, 2006), we conclude that the function $r(t)$ is non-increasing, which implies $r(t)\leq 0$ for all $t\in\mathbb{R}_+$.

Next, since the function $p(t)$ is continuous, the derivative $\dot{q}(t)$ of the function $q(t)$ exists and can be upper-bounded as follows:

$$\begin{aligned}\dot{q}(t) &\overset{(a)}{=} (\mathcal{M}_0+\mathcal{M}_1 p(t))\|x'-x\|_2\\ &\leq (\mathcal{M}_0+\mathcal{M}_1[p(0)+|p(t)-p(0)|])\|x'-x\|_2\\ &\overset{(b)}{\leq} (\mathcal{M}_0+\mathcal{M}_1[p(0)+q(t)])\|x'-x\|_2,\end{aligned}$$

where (a) uses the definition of $q(t)$; (b) uses the inequality $r(t)\leq 0$. After rearranging, we obtain the following inequality:

$$\dot{l}(t) \leq \mathcal{M}_1\|x'-x\|_2,$$

where $l(t) = \ln[\mathcal{M}_0 + \mathcal{M}_1[p(0) + q(t)]]$. By integrating the inequality, we obtain

$$\ln[\mathcal{M}_0 + \mathcal{M}_1[p(0) + q(1)]] - \ln[\mathcal{M}_0 + \mathcal{M}_1[p(0) + q(0)]] \leq \mathcal{M}_1\|x' - x\|_2.$$

By taking the exponent of both sides and rearranging, we obtain

$$\begin{aligned}\mathcal{M}_0 + \mathcal{M}_1[p(0) + q(1)] &\leq [\mathcal{M}_0 + \mathcal{M}_1[p(0) + q(0)]] \exp(\mathcal{M}_1\|x' - x\|_2) \\ &\overset{\text{(a)}}{=} [\mathcal{M}_0 + \mathcal{M}_1 p(0)] \exp(\mathcal{M}_1\|x' - x\|_2)\end{aligned}$$

where (a) uses the definition of $q(t)$. After rearranging, we obtain the following

$$\begin{aligned}q(1) &\leq [\mathcal{M}_0 + \mathcal{M}_1 p(0)] \tfrac{1}{\mathcal{M}_1}[\exp(\mathcal{M}_1\|x' - x\|_2) - 1] \\ &\overset{\text{(a)}}{\leq} [\mathcal{M}_0 + \mathcal{M}_1 p(0)] \exp(\mathcal{M}_1\|x' - x\|_2)\|x' - x\|_2,\end{aligned}$$

where (a) uses the convexity of the exponent. Finally, we use the inequality $r(1) \leq 0$. □

F Proofs for Section 3

F.1 Proof of Theorem 1

We define $\eta = \eta_k$ and proceed with the following inequality:

$$\begin{aligned}
\tfrac{1}{2}\|x_{k+1} - x^*\|_2^2 &\overset{(a)}{=} \tfrac{1}{2}\|\text{proj}_{Q_\mathcal{R}}(x_k - \eta\nabla f(x_k)) - \text{proj}_{Q_\mathcal{R}}(x^*)\|_2^2 \\
&\overset{(b)}{\leq} \tfrac{1}{2}\|x_k - x^* - \eta\nabla f(x_k)\|_2^2 \\
&= \tfrac{1}{2}\|x_k - x^*\|_2^2 - \eta\langle\nabla f(x_k), x_k - x^*\rangle + \tfrac{1}{2}\eta^2\|\nabla f(x_k)\|_2^2 \\
&\overset{(c)}{\leq} \tfrac{1}{2}\|x_k - x^*\|_2^2 - \eta[f(x_k) - f^*] + \tfrac{1}{2}\eta^2\|\nabla f(x_k)\|_2^2 \\
&\overset{(d)}{\leq} \tfrac{1}{2}\|x_k - x^*\|_2^2 - \eta[f(x_k) - f^*] + \tfrac{1}{2}\eta^2(\mathcal{M}_0 + \mathcal{M}_1[f(x_k) - f^*])^2 \\
&\overset{(e)}{\leq} \tfrac{1}{2}\|x_k - x^*\|_2^2 - \eta[f(x_k) - f^*] + \eta^2\mathcal{M}_0^2 + \eta^2\mathcal{M}_1^2\mathcal{F}[f(x_k) - f^*] \\
&\leq \tfrac{1}{2}\|x_k - x^*\|_2^2 - \tfrac{1}{2}\eta[f(x_k) - f^*] + \eta^2\mathcal{M}_0^2
\end{aligned}$$

where (a) uses the iterations in eq. (8) and the fact that $x^* \in Q_\mathcal{R}$; (b) uses the non-expansiveness of the projection; (c) uses the convexity of the function $f(x)$; (d) uses Assumption 1; (e) uses the definitions in eq. (6) and the fact that $\|x_k\|_2 \leq \mathcal{R}$. Next, we can obtain the following:

$$\begin{aligned}
f(\overline{x}_K) - f^* &\overset{(a)}{\leq} \frac{1}{K+1}\sum_{k=0}^{K}[f(x_k) - f^*] \overset{(b)}{\leq} \frac{\mathcal{R}^2}{\eta(K+1)} + 2\eta\mathcal{M}_0^2 \\
&\overset{(c)}{\leq} \mathcal{O}\Big[\frac{\mathcal{R}\mathcal{M}_0}{\sqrt{K+1}} + \frac{[\mathcal{R}\mathcal{M}_1]^2\mathcal{F}}{K+1}\Big] \overset{(d)}{\leq} \mathcal{O}[\epsilon],
\end{aligned} \tag{55}$$

where (a) uses the convexity of the function $f(x)$ and the definition of $\overline{x}_K$; (b) uses the inequality above and the fact that $\|x^*\|_2 \leq \mathcal{R}$ and $x_0 = 0$; (c) and (d) use the definitions in eq. (9). □

F.2 Proof of Theorem 2

We start with the following inequality:

$$\begin{aligned}
\tfrac{1}{2}\|x_{k+1} - x^*\|_2^2 &\overset{(a)}{=} \tfrac{1}{2}\|\text{proj}_{Q_\mathcal{R}}(x_k - \eta_k\nabla f(x_k)) - \text{proj}_{Q_\mathcal{R}}(x^*)\|_2^2 \\
&\overset{(b)}{\leq} \tfrac{1}{2}\|x_k - x^* - \eta_k\nabla f(x_k)\|_2^2 \\
&= \tfrac{1}{2}\|x_k - x^*\|_2^2 - \eta_k\langle\nabla f(x_k), x_k - x^*\rangle + \tfrac{1}{2}\eta_k^2\|\nabla f(x_k)\|_2^2 \\
&\overset{(c)}{\leq} \tfrac{1}{2}\|x_k - x^*\|_2^2 - \eta_k[f(x_k) - f^*] + \tfrac{1}{2}\eta_k^2\|\nabla f(x_k)\|_2^2,
\end{aligned}$$

where (a) uses the iterations in eq. (8) and the fact that $x^* \in Q_\mathcal{R}$; (b) uses the non-expansiveness of the projection; (c) uses the convexity of the function $f(x)$. Further, for Option A, we proceed as follows:

$$\begin{aligned}
\tfrac{1}{2}\|x_{k+1} - x^*\|_2^2 &\overset{(a)}{\leq} \tfrac{1}{2}\|x_k - x^*\|_2^2 - \frac{\mathcal{R}[f(x_k) - f^*]}{\sqrt{K+1}\|\nabla f(x_k)\|_2} + \frac{\mathcal{R}^2}{2[K+1]} \\
&\overset{(b)}{\leq} \tfrac{1}{2}\|x_k - x^*\|_2^2 - \frac{\mathcal{R}[f(x_k) - f^*]}{\sqrt{K+1}(\mathcal{M}_0 + \mathcal{M}_1[f(x_k) - f^*])} + \frac{\mathcal{R}^2}{2[K+1]} \\
&\leq \tfrac{1}{2}\|x_k - x^*\|_2^2 - \min\Big\{\frac{\mathcal{R}[f(x_k) - f^*]}{2\sqrt{K+1}\mathcal{M}_0}, \frac{\mathcal{R}}{2\sqrt{K+1}\mathcal{M}_1}\Big\} + \frac{\mathcal{R}^2}{2[K+1]},
\end{aligned}$$

where (a) uses Option A in eq. (11); (b) uses Assumption 1. After summing these inequalities and rearranging, we obtain the following:

$$\min\Big\{\min_{k\in\{0,\ldots,K\}}[f(x_k) - f^*], \frac{\mathcal{M}_0}{\mathcal{M}_1}\Big\} \leq \frac{2\mathcal{R}\mathcal{M}_0}{\sqrt{K+1}}.$$

From the choice of K in eq. (12), we have $\frac{\mathcal{M}_0}{\mathcal{M}_1} > \frac{2\mathcal{R}\mathcal{M}_0}{\sqrt{K+1}}$. Hence, the minimum in the inequality above is attained in the first term, which implies

$$\min_{k\in\{0,\ldots,K\}}[f(x_k) - f^*] \leq \frac{2\mathcal{R}\mathcal{M}_0}{\sqrt{K+1}} \leq \epsilon.$$

Finally, for Option B, we proceed as follows:

$$\begin{aligned}\tfrac{1}{2}\|x_{k+1}-x^*\|_2^2 &\overset{(a)}{\leq} \tfrac{1}{2}\|x_k-x^*\|_2^2 - \frac{[f(x_k)-f^*]^2}{2\|\nabla f(x_k)\|_2^2}\\ &\overset{(b)}{\leq} \tfrac{1}{2}\|x_k-x^*\|_2^2 - \frac{[f(x_k)-f^*]^2}{2[\mathcal{M}_0+\mathcal{M}_1[f(x_k)-f^*]]^2}\\ &\leq \tfrac{1}{2}\|x_k-x^*\|_2^2 - \min\Big\{\frac{[f(x_k)-f^*]^2}{8\mathcal{M}_0^2}, \frac{1}{8\mathcal{M}_1^2}\Big\},\end{aligned}$$

where (a) uses Option B in eq. (11); (b) uses Assumption 1. After summing these inequalities and rearranging, we obtain the following:

$$\min\Big\{\min_{k\in\{0,\dots,K\}}[f(x_k)-f^*], \frac{\mathcal{M}_0}{\mathcal{M}_1}\Big\}^2 \leq \frac{4[\mathcal{R}\mathcal{M}_0]^2}{K+1}.$$

The rest of the proof is identical to Option A. □

F.3 Proof of Theorem 3

We define $\eta = \eta_k$ and proceed with the following inequality:

$$\begin{aligned}\mathbb{E}\big[\tfrac{1}{2}\|x_{k+1}-x^*\|_2^2\big] &\overset{(a)}{=} \mathbb{E}\big[\tfrac{1}{2}\|\mathrm{proj}_{Q_\mathcal{R}}(x_k-\eta\nabla_{\xi_k}f(x_k)) - \mathrm{proj}_{Q_\mathcal{R}}(x^*)\|_2^2\big]\\ &\overset{(b)}{\leq} \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*-\eta\nabla_{\xi_k}f(x_k)\|_2^2\big]\\ &= \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*\|_2^2 - \eta\langle\nabla_{\xi_k}f(x_k), x_k-x^*\rangle + \tfrac{1}{2}\eta^2\|\nabla_{\xi_k}f(x_k)\|_2^2\big]\\ &\overset{(c)}{\leq} \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*\|_2^2 - \eta[f(x_k)-f^*] + \tfrac{1}{2}\eta^2\|\nabla_{\xi_k}f(x_k)\|_2^2\big]\\ &\overset{(d)}{\leq} \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*\|_2^2 - \eta[f(x_k)-f^*] + \eta^2\|u_{\xi_k}(x_k)\|_2^2 + \eta^2\|v_{\xi_k}(x_k)\|_2^2\big]\\ &\overset{(e)}{\leq} \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*\|_2^2 - \eta[f(x_k)-f^*] + \eta^2\mathcal{G}_0^2 + \eta^2\mathcal{G}_1^2[f(x_k)-f^*]^2\big]\\ &\overset{(f)}{\leq} \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*\|_2^2 - \eta[f(x_k)-f^*] + \eta^2\mathcal{G}_0^2 + \eta^2\mathcal{G}_1^2\mathcal{F}[f(x_k)-f^*]\big]\\ &\overset{(g)}{\leq} \mathbb{E}\big[\tfrac{1}{2}\|x_k-x^*\|_2^2 - \tfrac{1}{2}\eta[f(x_k)-f^*] + \eta^2\mathcal{G}_0^2\big],\end{aligned}$$

where (a) uses the iterations in eq. (13) and the fact that $x^* \in Q_\mathcal{R}$; (b) uses the non-expansiveness of the projection; (c) uses the unbiasedness property in Assumption 2 and the convexity of the function $f(x)$; (d) and (e) use Assumption 2; (f) uses the definition in eq. (6) and the fact that $\|x_k\|_2 \leq \mathcal{R}$; (g) uses the definition in eq. (15). After summing these inequalities, we obtain

$$\begin{aligned}\mathbb{E}[f(\overline{x}_K)-f^*] &\overset{(a)}{\leq} \frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f^*] \overset{(b)}{\leq} \frac{\mathcal{R}^2}{\eta(K+1)} + 2\eta\mathcal{G}_0^2\\ &\overset{(c)}{\leq} \mathcal{O}\Big[\frac{\mathcal{R}\mathcal{G}_0}{\sqrt{K+1}} + \frac{[\mathcal{R}\mathcal{G}_1]^2\mathcal{F}}{K+1}\Big] \overset{(d)}{\leq} \mathcal{O}[\epsilon],\end{aligned}$$

where (a) uses the convexity of the function $f(x)$ and the definition of $\overline{x}_K$; (b) uses the inequality above and the fact that $\|x^*\|_2 \leq \mathcal{R}$ and $x_0 = 0$; (c) uses the definitions in eq. (15); (d) uses the definition in eq. (15), Remark 1, and the inequality in eq. (7). □

F.4 Proof of Theorem 4

Using the definition of $\overline{x}_K$ and the convexity of $f(x)$, we can upper-bound $\mathbb{E}[f(\overline{x}_K)-f^*]$ as follows:

$$\mathbb{E}[f(\overline{x}_K)-f^*] \leq \frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f^*].$$

Next, we can upper-bound $\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f^*]$ as follows:

$$\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f^*] \overset{(a)}{\leq} \sum_{k=0}^{K}\mathbb{E}[\langle\nabla f(x_k), x_k-x^*\rangle]$$

$$
\begin{aligned}
&\overset{(b)}{=} \mathbb{E}\left[\sum_{k=0}^{K}\langle \nabla_{\xi_k} f(x_k), x_k - x^*\rangle\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\mathcal{R}\mathbb{E}\left[\sqrt{\textstyle\sum_{k=0}^{K}\|\nabla_{\xi_k} f(x_k)\|_2^2}\right]\right] \\
&\overset{(d)}{\leq} \mathcal{O}\left[\mathcal{R}\mathbb{E}\left[\sqrt{\textstyle\sum_{k=0}^{K}\|u_{\xi_k}(x_k) + v_{\xi_k}(x_k)\|_2^2}\right]\right] \\
&\leq \mathcal{O}\left[\mathcal{R}\mathbb{E}\left[\sqrt{\textstyle\sum_{k=0}^{K}\|u_{\xi_k}(x_k)\|_2^2} + \sqrt{\textstyle\sum_{k=0}^{K}\|v_{\xi_k}(x_k)\|_2^2}\right]\right] \\
&\overset{(e)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{K+1} + \mathcal{R}\mathcal{G}_1\mathbb{E}\left[\sqrt{\textstyle\sum_{k=0}^{K}[f(x_k) - f^*]^2}\right]\right] \\
&\overset{(f)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{K+1} + \mathcal{R}\mathcal{G}_1\mathbb{E}\left[\sqrt{\textstyle\sum_{k=0}^{K}\mathcal{F}[f(x_k) - f^*]}\right]\right] \\
&\overset{(g)}{\leq} \tfrac{1}{2}\sum_{k=0}^{K}\mathbb{E}[f(x_k) - f^*] + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{K+1} + [\mathcal{R}\mathcal{G}_1]^2\mathcal{F}\right],
\end{aligned}
$$

where (a) uses the convexity of the function $f(x)$; (b) uses the unbiasedness in Assumption 2; (c) uses the standard upper-bound on the regret for AdaGrad-Norm (Streeter & McMahan, 2010, Theorem 2); (d) uses Assumption 2; (e) uses Assumption 2, Jensen's inequality, and the concavity of the square root; (f) uses the definition in eq. (6) and the fact that $\|x_k\|_2 \leq \mathcal{R}$; (g) uses Young's inequality. After rearranging, we obtain the following:

$$
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k) - f^*] \leq \mathcal{O}\left[\frac{\mathcal{R}\mathcal{G}_0}{\sqrt{K+1}} + \frac{[\mathcal{R}\mathcal{G}_1]^2\mathcal{F}}{K+1}\right] \overset{(a)}{\leq} \mathcal{O}[\epsilon],
$$

where (a) uses the definition in eq. (16), Remark 1, and the inequality in eq. (7). □

G Proofs for Section 4

G.1 Technical Lemma

In this section, we present a technical lemma that will be used in the proofs of Theorems 5, 6 and 7.

Lemma 10. *Let $p \in (0, 1/2]$, and let $\{\Delta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ and $\{\delta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ be two sequences of non-negative numbers satisfying the following condition for all $k \in \mathbb{N}$:*

$$\pi_k\Delta_k \leq \delta_k + p\textstyle\sum_{i=1}^k \alpha_i\pi_i\Delta_i, \tag{56}$$

where α_k, π_k satisfy eq. (17). Then, the following inequalities hold for all $k \in \mathbb{N}$:

$$\textstyle\sum_{i=1}^k \alpha_i\pi_i\Delta_i \leq \sum_{i=2}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p + \alpha_1\delta_1 \min\{[\pi_k/\pi_0]^p, 2[\pi_k/\pi_1]^p\}. \tag{57}$$

Proof. We define $\tau_k \geq 0$ for $k \in \mathbb{N}$ as follows:

$$\tau_k = \textstyle\sum_{i=1}^k \alpha_i\pi_i\Delta_i.$$

We can upper-bound τ_k as follows:

$$\tau_k \overset{\text{(a)}}{=} \alpha_k\pi_k\Delta_k + \tau_{k-1} \overset{\text{(b)}}{\leq} \alpha_k\delta_k + p\alpha_k\tau_k + \tau_{k-1} \overset{\text{(c)}}{\leq} \alpha_k\delta_k + [1-(1-\alpha_k)^p]\tau_k + \tau_{k-1},$$

where (a) uses the definition of τ_k; (b) uses the assumption of Lemma 10; (c) uses the inequality $pt \leq 1-(1-t)^p$ for $t \leq 1$ which is implied by the concavity of the function $t \mapsto t^p$. After rearranging, we obtain the following:

$$\alpha_k\delta_k + \tau_{k-1} \geq (1-\alpha_k)^p\tau_k \overset{\text{(a)}}{=} [\pi_{k-1}/\pi_k]^p\tau_k$$

where (a) uses the definition in eq. (17). After rearranging one more time, we obtain the following:

$$[1/\pi_k]^p\tau_k \leq [1/\pi_{k-1}]^p[\tau_{k-1} + \alpha_k\delta_k].$$

In the case where $\pi_0 \neq 0$, we can upper-bound τ_k as follows:

$$\tau_k \leq \textstyle\sum_{i=1}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p.$$

In the case $\pi_0 = 0$, we can upper-bound τ_k as follows:

$$\begin{aligned}
\tau_k &\leq \textstyle\sum_{i=2}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p + \tau_1[\pi_k/\pi_1]^p \\
&\overset{\text{(a)}}{=} \textstyle\sum_{i=2}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p + \alpha_1\pi_1\Delta_1[\pi_k/\pi_1]^p \\
&\overset{\text{(b)}}{\leq} \textstyle\sum_{i=2}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p + 2\alpha_1\delta_1[\pi_k/\pi_1]^p,
\end{aligned}$$

where (a) uses the definition of τ_k; (b) uses the fact that $p\alpha_1 \leq 1/2$ and the assumption of Lemma 10. □

G.2 Proof of Lemma 2

We can lower-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_K] &\overset{\text{(a)}}{\geq} \sum_{k=1}^K \mathbb{E}[\langle \nabla_{\xi_k} f_k(z_k), z_k - x^*\rangle] \\
&\overset{\text{(b)}}{=} \sum_{k=1}^K \mathbb{E}[\langle \nabla f_k(z_k), z_k - x^*\rangle] \\
&\overset{\text{(c)}}{\geq} \sum_{k=1}^K \mathbb{E}[f_k(z_k) - f_k(x^*)] \\
&\overset{\text{(d)}}{=} \sum_{k=1}^K \pi_k\mathbb{E}[f(\alpha_k z_k + (1-\alpha_k)x_{k-1}) - f(\alpha_k x^* + (1-\alpha_k)x_{k-1})] \\
&\overset{\text{(e)}}{=} \sum_{k=1}^K \pi_k\mathbb{E}[f(x_k) - f(\alpha_k x^* + (1-\alpha_k)x_{k-1})]
\end{aligned}$$

$$\overset{(f)}{\geq} \sum_{k=1}^{K} (\pi_k \mathbb{E}[f(x_k) - f^*] - (1-\alpha_k)\pi_k \mathbb{E}[f(x_{k-1}) - f^*])$$
$$\overset{(g)}{=} \sum_{k=1}^{K} (\pi_k \mathbb{E}[f(x_k) - f^*] - \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*])$$
$$= \pi_K \mathbb{E}[f(x_K) - f^*] - \pi_0 [f(x_0) - f^*],$$

where (a) uses the definition in eq. (19); (b) uses the unbiasedness in Assumption 2, the fact that ξ_k is independent of z_k and x_{k-1}, and the definition in eq. (17); (c) uses the convexity of $f(x)$ and the definition in eq. (17); (d) and (g) use the definitions in eq. (17); (e) uses the iterations in eq. (18); (f) use the convexity of $f(x)$. Rearranging concludes the proof. □

G.3 Proof of Lemma 3

We can upper-bound $\mathbb{E}[\text{Reg}_K]$ as follows:

$$\mathbb{E}[\text{Reg}_K] \overset{(a)}{\leq} \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \|\nabla_{\xi_k} f_k(z_k)\|_2^2}\right]$$
$$\overset{(b)}{=} \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 \|\nabla_{\xi_k} f(x_k)\|_2^2}\right]$$
$$\overset{(c)}{\leq} \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 \|u_{\xi_k}(x_k) + v_{\xi_k}(x_k)\|_2^2}\right]$$
$$\leq \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 \|u_{\xi_k}(x_k)\|_2^2} + \mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 \|v_{\xi_k}(x_k)\|_2^2}\right]$$
$$\overset{(d)}{\leq} \mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 \mathbb{E}[\|u_{\xi_k}(x_k)\|_2^2]} + \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 \|v_{\xi_k}(x_k)\|_2^2}\right]$$
$$\overset{(e)}{\leq} \mathcal{C}\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2} + \mathbb{E}\left[\mathcal{C}\mathcal{R}\mathcal{G}_1\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 [f(x_k) - f^*]^2}\right],$$

where (a) uses Assumption 3 and the definition in eq. (19); (b) uses eq. (18) and the definition in eq. (17); (c) and (e) use Assumption 2; (d) uses Jensen's inequality and the concavity of the square root. □

G.4 Proof of Theorem 5

From the definition in eq. (17) and the definition $\alpha_k = 1/k$, it follows that $\pi_k = k$. Next, let $p \in (0, 1/2]$ and let $\Delta_k = \mathbb{E}[f(x_k) - f^*]$. Then, we can upper-bound $\pi_k \Delta_k = k\Delta_k$ as follows:

$$\pi_k \Delta_k \overset{(a)}{\leq} \mathbb{E}[\text{Reg}_k]$$
$$\overset{(b)}{\leq} \mathcal{C}\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{i=1}^{k} \alpha_i^2 \pi_i^2} + \mathbb{E}\left[\mathcal{C}\mathcal{R}\mathcal{G}_1\sqrt{\textstyle\sum_{i=1}^{k} \alpha_i^2 \pi_i^2 [f(x_i) - f^*]^2}\right]$$
$$\overset{(c)}{=} \mathcal{C}\mathcal{R}\mathcal{G}_0\sqrt{k} + \mathbb{E}\left[\mathcal{C}\mathcal{R}\mathcal{G}_1\sqrt{\textstyle\sum_{i=1}^{k} \alpha_i \pi_i [f(x_i) - f^*]^2}\right]$$
$$\overset{(d)}{\leq} \mathcal{C}\mathcal{R}\mathcal{G}_0\sqrt{k} + \mathcal{C}\mathcal{R}\mathcal{G}_1\sqrt{\textstyle\sum_{i=1}^{k} \mathcal{F}\alpha_i \pi_i \Delta_i}$$
$$\overset{(e)}{\leq} \mathcal{C}\mathcal{R}\mathcal{G}_0\sqrt{k} + \tfrac{1}{4p}[\mathcal{C}\mathcal{R}\mathcal{G}_1]^2\mathcal{F} + p\textstyle\sum_{i=1}^{k} \alpha_i \pi_i \Delta_i$$

where (a) uses Lemma 2; (b) uses Lemma 3; (c) uses the fact that $\alpha_k \pi_k = 1$; (d) uses the definition of $\mathcal{F}$, Jensen's inequality, and the definition of Δ_i; (e) uses Young's inequality. Now, we can define $\delta_k = \mathcal{C}\mathcal{R}\mathcal{G}_0\sqrt{k} + \frac{1}{4p}[\mathcal{C}\mathcal{R}\mathcal{G}_1]^2\mathcal{F}$ and $p = \min\left\{\frac{1}{\ln K}, \frac{1}{4}\right\}$, and obtain the following inequality:

$$\textstyle\sum_{i=1}^{k} \alpha_i \pi_i \Delta_i$$
$$\overset{(a)}{\leq} \textstyle\sum_{i=2}^{k} \alpha_i \delta_i \left[\frac{\pi_k}{\pi_{i-1}}\right]^p + 2\alpha_1 \delta_1 \left[\frac{\pi_k}{\pi_1}\right]^p$$

$$\overset{(b)}{=} \sum_{i=2}^{k} \alpha_i \left[\frac{\pi_k}{\pi_{i-1}}\right]^p \left[\mathcal{CRG}_0\sqrt{i} + \frac{1}{4p}[\mathcal{CRG}_1]^2\mathcal{F}\right] + 2\alpha_1\left[\frac{\pi_k}{\pi_1}\right]^p \left[\mathcal{CRG}_0 + \frac{[\mathcal{CRG}_1]^2\mathcal{F}}{4p}\right]$$
$$\overset{(c)}{=} \sum_{i=2}^{k} \frac{k^p}{i(i-1)^p} \left[\mathcal{CRG}_0\sqrt{i} + \frac{1}{4p}[\mathcal{CRG}_1]^2\mathcal{F}\right] + 2k^p \left[\mathcal{CRG}_0 + \frac{[\mathcal{CRG}_1]^2\mathcal{F}}{4p}\right]$$
$$\leq k^p \sum_{i=1}^{k-1} \left[\frac{\mathcal{CRG}_0}{i^{p+1/2}} + \frac{[\mathcal{CRG}_1]^2\mathcal{F}}{4pi^{p+1}}\right] + 2k^p \left[\mathcal{CRG}_0 + \frac{[\mathcal{CRG}_1]^2\mathcal{F}}{4p}\right]$$
$$\overset{(d)}{\leq} k^p \left[\mathcal{CRG}_0 \left[1 + \frac{k^{1/2-p}-1}{1/2-p}\right] + \frac{(p+1)[\mathcal{CRG}_1]^2\mathcal{F}}{4p^2} + 2\mathcal{CRG}_0 + \frac{[\mathcal{CRG}_1]^2\mathcal{F}}{2p}\right]$$
$$\overset{(e)}{\leq} \left[4\mathcal{CRG}_0\sqrt{k} + \frac{[\mathcal{CRG}_1]^2\mathcal{F}k^p}{2p^2}\right]$$

where (a) uses Lemma 10; (b) uses the definition of δ_i; (c) uses the fact that $\alpha_i = 1/i$ and $\pi_i = i$; (d) uses the fact that $p + 1 > 1$ and $p + 1/2 \in (0, 1)$; (e) uses the fact that $p \leq 1/4$. Next, we upper-bound Δ_K as follows:

$$\Delta_K \overset{(a)}{\leq} \frac{\mathcal{CRG}_0}{\sqrt{K}} + \frac{\mathcal{CRG}_1}{K}\sqrt{\sum_{k=1}^{K} \mathcal{F}\alpha_k\pi_k\Delta_k}$$
$$\overset{(b)}{\leq} \frac{\mathcal{CRG}_0}{\sqrt{K}} + \frac{\mathcal{CRG}_1}{K}\sqrt{4\mathcal{CRG}_0\mathcal{F}K^{1/2} + \frac{[\mathcal{CRG}_1\mathcal{F}]^2K^p}{2p^2}}$$
$$\overset{(c)}{\leq} \frac{\mathcal{CRG}_0}{\sqrt{K}} + \frac{\mathcal{CRG}_1}{K}\sqrt{4\mathcal{CRG}_0\mathcal{F}K^{1/2} + \frac{[\mathcal{CRG}_1\mathcal{F}]^2e}{2p^2}}$$
$$\leq \frac{\mathcal{CRG}_0}{\sqrt{K}} + \frac{2\sqrt{\mathcal{CRG}_0\mathcal{F}}\mathcal{CRG}_1}{K^{3/4}} + \frac{\sqrt{e}[\mathcal{CRG}_1]^2\mathcal{F}}{\sqrt{2}pK}$$
$$\overset{(d)}{\leq} \mathcal{O}\left[\frac{\mathcal{RG}_0}{\sqrt{K}} + \frac{[1+\ln K][\mathcal{RG}_1]^2\mathcal{F}}{K}\right],$$

where (a) uses the previously obtained upper bound on $\pi_K\Delta_K$ and the fact that $\pi_K = K$; (b) uses the previously obtained upper bound on $\sum_{k=1}^{K}\alpha_k\pi_k\Delta_k$; (c) uses the fact that $p \leq 1/\ln K$; (d) uses Young's inequality and the choice of p. Finally, we define $\kappa = e + [\mathcal{RG}_0/\epsilon] + [\mathcal{RG}_1]^2 \cdot [\mathcal{F}/\epsilon]$ and upper-bound Δ_K as follows:

$$\Delta_K \leq \mathcal{O}\left[\frac{\mathcal{RG}_0}{\sqrt{K}} + \frac{[1+\ln K][\mathcal{RG}_1]^2\mathcal{F}}{K}\right]$$
$$\overset{(a)}{\leq} \mathcal{O}\left[\epsilon + \epsilon\ln^{-1}[\kappa]\left[1 + \ln\left[1 + [\mathcal{RG}_0/\epsilon]^2 + [\mathcal{RG}_1]^2 \cdot [\mathcal{F}/\epsilon] \cdot \ln[\kappa]\right]\right]\right]$$
$$\leq \mathcal{O}\left[\epsilon + \epsilon\ln^{-1}[\kappa]\left[1 + \ln\left[\kappa^2\ln[\kappa]\right]\right]\right]$$
$$\leq \mathcal{O}\left[\epsilon + \epsilon\ln^{-1}[\kappa](1 + 2\ln[\kappa] + \ln[\ln[\kappa]])\right]$$
$$\leq \mathcal{O}\left[\epsilon + \epsilon\ln^{-1}[\kappa](1 + 3\ln[\kappa])\right]$$
$$\leq \mathcal{O}[\epsilon],$$

where (a) uses the definition in eq. (23). □

G.5 Proof of Lemma 4

First, let $J = \lfloor K/T \rfloor \geq 1$ and define $\mathcal{K}_s \subset \mathbb{N}$ for $s \in \{1, \ldots, J\}$ as follows:

$$\mathcal{K}_s = \begin{cases} \{(s-1)T + 1, \ldots, sT\} & s \in \{1, \ldots, J-1\} \\ \{(s-1)T + 1, \ldots, K\} & s = J \end{cases},$$

and let $k_s = \min \mathcal{K}_s$. It is easy to observe that $\sqcup_{s=1}^{J}\mathcal{K}_s = \{1, \ldots, K\}$ and $T \leq |\mathcal{K}_s| \leq 2T - 1$. Further, we can upper-bound $\|x_k - x_{k_s}\|_2$ for $k \in \mathcal{K}_s$ as

$$\|x_k - x_{k_s}\|_2 \overset{(a)}{\leq} \sum_{i=k_s}^{k-1}\|x_{i+1} - x_i\|_2 \overset{(b)}{\leq} \sum_{i=k_s}^{k-1}\alpha\|z_{i+1} - x_i\|_2 \overset{(c)}{\leq} \sum_{i=k_s}^{k-1} 2\alpha\mathcal{R} \overset{(d)}{\leq} \mathcal{O}[\alpha T\mathcal{R}] \overset{(e)}{\leq} \mathcal{O}[1/\mathcal{G}_1],$$

where (a) uses the triangle inequality; (b) uses the definition in eq. (18) and the assumption in eq. (25); (c) uses the fact that $z_{i+1}, x_i \in Q_{\mathcal{R}}$; (d) use the fact that $k, k_s \in \mathcal{K}_s$ and the properties of $\mathcal{K}_s$; (e) uses the assumptions in eq. (25). Next, we can upper-bound π_k for $k \in \mathcal{K}_s$ as follows:

$$\pi_k \overset{(a)}{\leq} \pi_{k_s}\prod_{i=k_s+1}^{k}(1-\alpha_i)^{-1} \overset{(b)}{\leq} \pi_{k_s}(1-1/T)^{1-2T} = \pi_{k_s}(1 + 1/(T-1))^{2T-1} \leq \mathcal{O}[\pi_{k_s}],$$

where (a) uses the definition in eq. (17); (b) uses the assumptions in eq. (25) and the fact that $k \in \mathcal{K}_s$. Next, we can obtain the following inequality:

$$\begin{aligned}
&\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2[f(x_k)-f^*]^2} = \sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k^2[f(x_k)-f^*]^2} \\
&\leq \sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k^2[f(x_{k_s})-f^*]^2} + \sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k^2[f(x_k)-f(x_{k_s})]^2} \\
&\overset{(a)}{\leq} \sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k^2[f(x_{k_s})-f^*]^2} + \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k^2([\mathcal{G}_0/\mathcal{G}_1]+[f(x_{k_s})-f^*])^2}\right] \\
&\leq \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k^2[f(x_{k_s})-f^*]^2} + [\mathcal{G}_0/\mathcal{G}_1]\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2}\right] \\
&\overset{(b)}{=} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha^2\pi_{k_s}^2[f(x_{k_s})-f^*]^2} + [\mathcal{G}_0/\mathcal{G}_1]\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2}\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\textstyle\sum_{s=1}^{J}\sqrt{T}\alpha\pi_{k_s}[f(x_{k_s})-f^*] + [\mathcal{G}_0/\mathcal{G}_1]\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2}\right],
\end{aligned}$$

where (a) uses Lemma 1, Remark 1, and the upper bound on $\|x_k - x_{k_s}\|_2$ for $k \in \mathcal{K}_s$ above; (b) uses the assumption in eq. (25) and the upper bound on π_k above; (c) uses the properties of $\mathcal{K}_s$. Furthermore, we can obtain the following inequality:

$$\begin{aligned}
\textstyle\sum_{s=1}^{J}\sqrt{T}\alpha\pi_{k_s}[f(x_{k_s})-f^*] &\overset{(a)}{\leq} \mathcal{O}\left[\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha\pi_{k_s}}{\sqrt{T}}[f(x_{k_s})-f(x_k)+f(x_k)-f^*]\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha\pi_{k_s}}{\sqrt{T}}(\alpha T\mathcal{R}\mathcal{G}_0+[f(x_k)-f^*])\right] \\
&\leq \mathcal{O}\left[\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha\pi_{k_s}}{\sqrt{T}}[f(x_k)-f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha^2\pi_{k_s}\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_k}{\sqrt{T}}[f(x_k)-f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\textstyle\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_k\right] \\
&\overset{(d)}{\leq} \mathcal{O}\left[\tfrac{1}{\sqrt{T}}\textstyle\sum_{k=1}^{K}\alpha_k\pi_k[f(x_k)-f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k\right],
\end{aligned}$$

where (a) use the properties of $\mathcal{K}_s$; (b) uses Lemma 1, Remark 1, and the upper bounds on $\|x_k - x_{k_s}\|_2$ for $k \in \mathcal{K}_s$ above; (c) uses the assumption in eq. (25) and the definition in eq. (17); (d) uses the properties of $\mathcal{K}_s$. Finally, we can upper-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows:

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_K] &\overset{(a)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathbb{E}\left[\mathcal{R}\mathcal{G}_1\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2[f(x_k)-f^*]^2}\right]\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathcal{R}\mathcal{G}_1\textstyle\sum_{s=1}^{J}\sqrt{T}\alpha_{k_s}\pi_{k_s}\mathbb{E}[f(x_{k_s})-f^*]\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \sqrt{T}\mathcal{R}^2\mathcal{G}_0\mathcal{G}_1\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k + \frac{\mathcal{R}\mathcal{G}_1}{\sqrt{T}}\textstyle\sum_{k=1}^{K}\alpha_k\pi_k\mathbb{E}[f(x_k)-f^*]\right] \\
&\overset{(d)}{\leq} \tfrac{1}{2}\textstyle\sum_{k=1}^{K}\alpha_k\pi_k\mathbb{E}[f(x_k)-f^*] + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + T\mathcal{R}\mathcal{G}_0\textstyle\sum_{k=1}^{K}\alpha_k^2\pi_k\right],
\end{aligned}$$

where (a) uses Lemma 3; (b) and (c) use the inequalities above; (d) uses the definition in eq. (25). □

G.6 Proof of Theorem 6

Let $\Delta_k = \mathbb{E}[f(x_k) - f^*]$, and let δ_k be defined as follows:

$$\delta_k = e\mathcal{F} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{i=1}^{k}\alpha_i^2\pi_i^2} + T\mathcal{R}\mathcal{G}_0\textstyle\sum_{i=1}^{k}\alpha_i^2\pi_i\right].$$

For $k \leq T-1$, we can upper-bound Δ_k as follows:

$$\pi_k\Delta_k \overset{(a)}{\leq} \pi_k\mathcal{F} \overset{(b)}{\leq} (1-1/T)^{1-T}\mathcal{F} \leq e\mathcal{F} \overset{(c)}{\leq} \delta_k \leq \delta_k + \tfrac{1}{2}\textstyle\sum_{i=1}^{k}\alpha_i\pi_i\Delta_i,$$

where (a) uses the definition in eq. (6) and the fact that $x_k \in Q_{\mathcal{R}}$; (b) uses the definitions in eqs. (17) and (27); (c) uses the definition of δ_k. For $k \geq T$, we can upper-bound Δ_k as follows:

$$\pi_k \Delta_k \overset{(a)}{\leq} \Delta_0 + \tfrac{1}{2}\textstyle\sum_{i=1}^{k} \alpha_i \pi_i \Delta_i + \mathcal{O}\Big[\mathcal{R}\mathcal{G}_0 \sqrt{\textstyle\sum_{i=1}^{k} \alpha_i^2 \pi_i^2} + T\mathcal{R}\mathcal{G}_0 \textstyle\sum_{i=1}^{k} \alpha_i^2 \pi_i\Big]$$
$$\overset{(b)}{\leq} \delta_k + \tfrac{1}{2}\textstyle\sum_{i=1}^{k} \alpha_i \pi_i \Delta_i,$$

where (a) uses Lemmas 2 and 4; (b) uses the definition in eq. (6) and the definition of δ_k. Hence, the conditions of Lemma 10 with $p = 1/2$ are satisfied. Further, we can upper-bound δ_k as follows:

$$\delta_k \overset{(a)}{=} e\mathcal{F} + \mathcal{O}\Big[\mathcal{R}\mathcal{G}_0 \alpha \pi_k \sqrt{\textstyle\sum_{i=0}^{k-1} (1-\alpha)^{2i}} + T\mathcal{R}\mathcal{G}_0 \alpha^2 \pi_k \textstyle\sum_{i=0}^{k-1} (1-\alpha)^i\Big]$$
$$\leq e\mathcal{F} + \mathcal{O}\big[\mathcal{R}\mathcal{G}_0 \sqrt{\alpha} \pi_k + T\mathcal{R}\mathcal{G}_0 \alpha \pi_k\big]$$
$$\overset{(b)}{\leq} e\mathcal{F} + \mathcal{O}[\epsilon \pi_k],$$

where (a) uses the definitions in eqs. (17) and (27); (b) uses the definition in eq. (27). Finally, we can upper-bound Δ_K as follows:

$$\Delta_K \overset{(a)}{\leq} \mathcal{O}\Big[\tfrac{\delta_K}{\pi_K} + \textstyle\sum_{k=1}^{K} \frac{\alpha_k \delta_k}{\sqrt{\pi_K \pi_{k-1}}}\Big]$$
$$\overset{(b)}{\leq} \mathcal{O}\Big[\tfrac{\mathcal{F}}{\pi_K} + \epsilon + \textstyle\sum_{k=1}^{K} \frac{\alpha_k \mathcal{F}}{\sqrt{\pi_K \pi_{k-1}}} + \sum_{k=1}^{K} \frac{\alpha_k \pi_k \epsilon}{\sqrt{\pi_K \pi_{k-1}}}\Big]$$
$$\overset{(c)}{\leq} \mathcal{O}\Big[\mathcal{F}[1-\alpha_1]^K + \epsilon + \alpha_1 \mathcal{F} \textstyle\sum_{k=0}^{K-1} [1-\alpha_1]^{[K+k]/2} + \alpha_1 \epsilon \sum_{k=0}^{K-1} [1-\alpha_1]^{[K-k]/2-1}\Big]$$
$$\leq \mathcal{O}\Big[\mathcal{F}[1-\alpha_1]^{K/2} + \epsilon\Big]$$
$$\overset{(d)}{\leq} \mathcal{O}[\epsilon],$$

where (a) uses Lemma 10 with $p = 1/2$; (b) uses the inequality above; (c) uses the definitions in eqs. (17) and (27); (d) uses the definition in eq. (27). □

G.7 Proof of Lemma 5

We define $J \in \mathbb{N}$, $k_s \in \mathbb{N}$ and $\mathcal{K}_s \subset \mathbb{N}$ for $s \in \{1, \ldots, J\}$ similarly to the proof of Lemma 4 in Section G.5. Consequently, it holds that $\pi_k \leq \mathcal{O}[\pi_{k_s}]$. Furthermore, for $k \in \mathcal{K}_s$, using the fact that $\alpha_k \leq \alpha_{k_s}$ due to the definition in eq. (29), we can upper-bound $\|x_k - x_{k_s}\|_2$ as follows:

$$\|x_k - x_{k_s}\|_2 \leq \mathcal{O}[\alpha_{k_s} T\mathcal{R}] \leq \mathcal{O}[1/\mathcal{G}_1].$$

Next, following the proof of Lemma 4 in Section G.5, we can obtain the following inequality:

$$\sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2 [f(x_k) - f^*]^2} \leq \mathcal{O}\Big[\sqrt{\textstyle\sum_{s=1}^{J} \sum_{k \in \mathcal{K}_s} \alpha_k^2 \pi_k^2 [f(x_{k_s}) - f^*]^2} + [\mathcal{G}_0/\mathcal{G}_1] \sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2}\Big]$$
$$\overset{(a)}{\leq} \mathcal{O}\Big[\textstyle\sum_{s=1}^{J} \sqrt{T} \alpha_{k_s} \pi_{k_s} [f(x_{k_s}) - f^*] + [\mathcal{G}_0/\mathcal{G}_1] \sqrt{\textstyle\sum_{k=1}^{K} \alpha_k^2 \pi_k^2}\Big],$$

where (a) uses the fact that $\alpha_k \leq \alpha_{k_s}$ and $\pi_k \leq \mathcal{O}[\pi_{k_s}]$, and the properties of $\mathcal{K}_s$. Furthermore, from the definition in eq. (29) it follows that for $s \leq J-1$ we have $\alpha_k = \alpha_{k_s}$ for all $k \in \mathcal{K}_s$, which implies $\alpha_{k_s} = \sum_{k \in \mathcal{K}_s} [\alpha_k / T]$. In addition, for $s = J$, we have

$$\alpha_{k_J} = \textstyle\sum_{k=(J-1)T+1}^{JT} [\alpha_k / T] \leq \sum_{k \in \mathcal{K}_J} [\alpha_k / T].$$

Using similar arguments, we can show that

$$\textstyle\sum_{k \in \mathcal{K}_s} \alpha_{k_s}^2 \leq \mathcal{O}\big[\sum_{k \in \mathcal{K}_s} \alpha_k^2\big] \quad \text{for all} \quad s \in \{1, \ldots, J\}.$$

Hence, we can obtain the following inequality:

$$\textstyle\sum_{s=1}^{J} \sqrt{T} \alpha_{k_s} \pi_{k_s} [f(x_{k_s}) - f^*] \leq \mathcal{O}\Big[\sum_{s=1}^{J} \sum_{k \in \mathcal{K}_s} \frac{\alpha_k \pi_{k_s}}{\sqrt{T}} [f(x_{k_s}) - f(x_k) + f(x_k) - f^*]\Big]$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}[\alpha_{k_s}T\mathcal{R}\mathcal{G}_0 + [f(x_k) - f^*]]\right] \\
&\leq \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}[f(x_k) - f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k\alpha_{k_s}\pi_{k_s}\right] \\
&\leq \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}[f(x_k) - f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}[\alpha_{k_s}^2 + \alpha_k^2]\pi_{k_s}\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}[f(x_k) - f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\alpha_k^2\pi_{k_s}\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\sum_{k=1}^{K}\frac{\alpha_k\pi_k}{\sqrt{T}}[f(x_k) - f^*] + \sqrt{T}\mathcal{R}\mathcal{G}_0\sum_{k=1}^{K}\alpha_k^2\pi_k\right],
\end{aligned}
$$

where (a) uses Lemma 1, Remark 1, and the upper bounds on $\|x_k - x_{k_s}\|_2$ for $k \in \mathcal{K}_s$ above; (b) uses the previously obtained inequalities; (c) uses the definition in eq. (17) and the properties of $\mathcal{K}_s$. Finally, we can upper-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows:

$$
\begin{aligned}
\mathbb{E}[\mathrm{Reg}_K] &\overset{(a)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathbb{E}\left[\mathcal{R}\mathcal{G}_1\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2[f(x_k) - f^*]^2}\right]\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathcal{R}\mathcal{G}_1\sum_{s=1}^{J}\sqrt{T}\alpha_{k_s}\pi_{k_s}\mathbb{E}[f(x_{k_s}) - f^*]\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \sqrt{T}\mathcal{R}^2\mathcal{G}_0\mathcal{G}_1\sum_{k=1}^{K}\alpha_k^2\pi_k + \frac{\mathcal{R}\mathcal{G}_1}{\sqrt{T}}\sum_{k=1}^{K}\alpha_k\pi_k\mathbb{E}[f(x_k) - f^*]\right] \\
&\overset{(d)}{\leq} \tfrac{1}{2}\sum_{k=1}^{K}\alpha_k\pi_k\mathbb{E}[f(x_k) - f^*] + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + T\mathcal{R}\mathcal{G}_0\sum_{k=1}^{K}\alpha_k^2\pi_k\right],
\end{aligned}
$$

where (a) uses Lemma 3; (b) and (c) use the inequalities above; (d) uses the definition in eq. (25). □

G.8 Proof of Theorem 7

First, observe that the statement of Lemma 5 holds for arbitrary $K \geq T$. Next, we define Δ_k and δ_k similarly to the proof of Theorem 6 in Section G.6. Consequently, using Lemmas 2 and 5 and using similar arguments to Section G.6, we can show that the following inequality holds for all $k \in \mathbb{N}$:

$$
\pi_k\Delta_k \leq \delta_k + \tfrac{1}{2}\sum_{i=1}^{k}\alpha_i\pi_i\Delta_i.
$$

Hence, the conditions of Lemma 10 with $p = 1/2$ are satisfied. Furthermore, for $k \leq ST$, we can upper-bound δ_k as follows:

$$
\begin{aligned}
\delta_k &\overset{(a)}{=} e\mathcal{F} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\alpha_1\pi_k\sqrt{\sum_{i=0}^{k-1}(1-\alpha_1)^{2i}} + T\mathcal{R}\mathcal{G}_0\alpha_1^2\pi_k\sum_{i=0}^{k-1}(1-\alpha_1)^i\right] \\
&\leq e\mathcal{F} + \mathcal{O}[\mathcal{R}\mathcal{G}_0\sqrt{\alpha_1}\pi_k + T\mathcal{R}\mathcal{G}_0\alpha_1\pi_k] \\
&\overset{(b)}{\leq} e\mathcal{F} + \mathcal{O}[[\mathcal{G}_0/\mathcal{G}_1]\pi_k],
\end{aligned}
$$

where (a) uses the definitions in eqs. (17) and (29); (b) uses the definition in eq. (29). For $k \geq ST+1$, we can upper-bound δ_k as follows:

$$
\begin{aligned}
\delta_k &\overset{(a)}{=} \delta_{ST} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\alpha_{ST+1}\pi_k\sqrt{\sum_{i=0}^{\infty}(1-\alpha_{ST+1})^{2i}} + T\mathcal{R}\mathcal{G}_0\alpha_{ST+1}^2\pi_k\sum_{i=0}^{\infty}(1-\alpha_{ST+1})^i\right] \\
&\leq \delta_{ST} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\alpha_{ST+1}}\pi_k + T\mathcal{R}\mathcal{G}_0\alpha_{ST+1}\pi_k\right] \\
&\overset{(b)}{\leq} \delta_{ST} + \mathcal{O}[\epsilon\pi_k],
\end{aligned}
$$

where (a) uses the definitions in eqs. (17) and (29); (b) uses the definition in eq. (29). Finally, we can upper-bound Δ_K as follows:

$$
\begin{aligned}
\Delta_K &\overset{(a)}{\leq} \mathcal{O}\left[\frac{\delta_K}{\pi_K} + \sum_{k=1}^{K}\frac{\alpha_k\delta_k}{\sqrt{\pi_K\pi_{k-1}}}\right] \\
&\overset{(b)}{=} \mathcal{O}\left[\frac{\delta_K}{\pi_K} + \sum_{k=1}^{ST}\frac{\alpha_k\delta_k}{\sqrt{\pi_K\pi_{k-1}}} + \sum_{k=ST+1}^{K}\frac{\alpha_k\delta_k}{\sqrt{\pi_K\pi_{k-1}}}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \mathcal{O}\left[\frac{\mathcal{F}}{\pi_K} + \frac{\mathcal{G}_0\pi_{ST}}{\mathcal{G}_1\pi_K} + \epsilon + \sum_{k=1}^{ST}\left[\frac{\alpha_k\mathcal{F}}{\sqrt{\pi_K\pi_{k-1}}} + \frac{\mathcal{G}_0\alpha_k\pi_k}{\mathcal{G}_1\sqrt{\pi_K\pi_{k-1}}}\right]\right] \\
&+ \mathcal{O}\left[\sum_{k=ST+1}^{K}\left[\frac{\alpha_k\mathcal{F}}{\sqrt{\pi_K\pi_{k-1}}} + \frac{\mathcal{G}_0\alpha_k\pi_{ST}}{\mathcal{G}_1\sqrt{\pi_K\pi_{k-1}}} + \frac{\epsilon\alpha_k\pi_k}{\sqrt{\pi_K\pi_{k-1}}}\right]\right] \\
&\overset{(d)}{\leq} \mathcal{O}\left[\frac{\mathcal{F}}{\pi_K} + \frac{\mathcal{G}_0\pi_{ST}}{\mathcal{G}_1\pi_K} + \epsilon + \frac{\mathcal{F}\alpha_1}{\sqrt{\pi_K}}\sum_{k=0}^{ST-1}[1-\alpha_1]^{k/2} + \frac{\mathcal{G}_0\alpha_1}{\mathcal{G}_1\sqrt{\pi_K}}\sum_{k=0}^{ST-1}[1-\alpha_1]^{-k/2-1}\right] \\
&+ \mathcal{O}\left[\left[\frac{\alpha_{ST+1}\mathcal{F}}{\sqrt{\pi_K\pi_{ST}}} + \frac{\mathcal{G}_0\alpha_{ST+1}\sqrt{\pi_{ST}}}{\mathcal{G}_1\sqrt{\pi_K}}\right]\sum_{k=0}^{K-ST-1}[1-\alpha_{ST+1}]^{k/2}\right] \\
&+ \mathcal{O}\left[\frac{\epsilon\alpha_{ST+1}\sqrt{\pi_{ST}}}{2\sqrt{\pi_K}}\sum_{k=0}^{K-ST-1}[1-\alpha_{ST+1}]^{-k/2-1}\right] \\
&\overset{(e)}{\leq} \mathcal{O}\left[\mathcal{F}[1-\alpha_1]^{ST/2}[1-\alpha_{ST+1}]^{[K-ST]/2} + [\mathcal{G}_0/\mathcal{G}_1][1-\alpha_{ST+1}]^{[K-ST]/2} + \epsilon\right] \\
&\overset{(f)}{\leq} \mathcal{O}[\epsilon],
\end{aligned}
$$

where (a) uses Lemma 10 with $p = 1/2$; (b) uses the fact that $K \geq TS + 1$; (c) uses the inequalities above; (d) and (e) use the definitions in eqs. (17) and (29); (f) uses the definitions in eq. (30). □

H PROOFS FOR SECTION A

H.1 PROOF OF THEOREM 8

In this proof, we consider one-dimensional example functions $f(x)$. The construction is inspired by the counterexamples for the generalized smooth optimization setting (Crawshaw et al., 2022; Zhang et al., 2019). We choose deterministic gradient oracle $\nabla_\xi f(x) = \nabla f(x)$, where $\nabla f(x) \in \partial f(x)$ is an arbitrary subgradient. We further consider the following two cases:

$$\text{Case I:} \quad \eta > \tfrac{2\mathcal{R}}{\mathcal{G}_0}\exp(-\tfrac{1}{4}\mathcal{R}\mathcal{G}_1), \qquad \text{Case II:} \quad \eta \le \tfrac{2\mathcal{R}}{\mathcal{G}_0}\exp(-\tfrac{1}{4}\mathcal{R}\mathcal{G}_1).$$

Case I. Consider the following example:

$$f(x) = \begin{cases} \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1|x - \frac{3}{4}\mathcal{R}|) - 1] & x \ge \frac{1}{2}\mathcal{R} \\ \mathcal{G}_0\exp(\frac{1}{4}\mathcal{R}\mathcal{G}_1)[\frac{1}{2}\mathcal{R} - x] + \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\frac{1}{4}\mathcal{R}\mathcal{G}_1) - 1] & x < \frac{1}{2}\mathcal{R} \end{cases}$$

It is easy to verify that the gradient $\nabla f(x)$ satisfies the following:

$$\nabla f(0) = \nabla f(-\mathcal{R}) = -\mathcal{G}_0\exp(\tfrac{1}{4}\mathcal{R}\mathcal{G}_1), \quad \nabla f(\mathcal{R}) = \mathcal{G}_0\exp(\tfrac{1}{4}\mathcal{R}\mathcal{G}_1).$$

Since $x_0 = \overline{x}_0 = 0$, using the lower bound on the stepsize η and eq. (13), for $K \in \mathbb{N}$, we get

$$x_K = (-1)^{K+1}\cdot\mathcal{R}, \quad \overline{x}_K = \begin{cases} 0 & K \equiv 0 \bmod 2 \\ \mathcal{R}/(K+1) & K \equiv 1 \bmod 2 \end{cases}.$$

Hence, for all $K \in \mathbb{N}$, it is easy to verify that $\overline{x}_K \le \frac{1}{2}\mathcal{R}$ and

$$\begin{aligned}\min\{f(x_K), f(\overline{x}_K)\} - f^* &\ge \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\tfrac{1}{4}\mathcal{R}\mathcal{G}_1) - 1] \\ &\overset{(a)}{\ge} \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\tfrac{1}{8}\mathcal{R}\mathcal{G}_1 + 1) - 1] \overset{(b)}{>} \tfrac{\mathcal{G}_0}{\mathcal{G}_1}\exp(\tfrac{1}{8}\mathcal{R}\mathcal{G}_1) \overset{(c)}{\ge} \epsilon,\end{aligned}$$

where (a), (b) and (c) use the assumptions in eq. (32). This concludes Case I.

Case II. Let $r = \max\left\{0, \frac{1}{\mathcal{G}_1}\ln\left[\frac{8\epsilon}{\mathcal{R}\mathcal{G}_0}\right]\right\}$, and consider the following example:

$$f(x) = \begin{cases} \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1|x - \frac{1}{2}\mathcal{R}|) - 1] & |x - \frac{1}{2}\mathcal{R}| \le r \\ \frac{8\epsilon}{\mathcal{R}}[|x - \frac{1}{2}\mathcal{R}| - r] + \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] & |x - \frac{1}{2}\mathcal{R}| > r \end{cases}.$$

One can verify that $f(x)$ is convex and $\nabla f(x)$ satisfies Assumption 2. Indeed, in the case $r = 0$, the convexity is obvious and $\frac{8\epsilon}{\mathcal{R}} \le \mathcal{G}_0$. In the case $r > 0$, we can show that $\nabla f(\frac{1}{2}\mathcal{R} \pm r) = \pm\frac{8\epsilon}{\mathcal{R}}$, and hence, the function $f(x)$ is differentiable at $x = \frac{1}{2}\mathcal{R} \pm r$. Furthermore, using the assumptions in eq. (32), we can show that $r \le \frac{1}{8}\mathcal{R}$. Hence, the following inequality holds:

$$f(\tfrac{1}{8}\mathcal{R}) = \tfrac{8\epsilon}{\mathcal{R}}\left[\tfrac{3}{8}\mathcal{R} - r\right] + \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] \ge 2\epsilon + \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] \ge 2\epsilon.$$

Consequently, we can show by induction that $x_K, \overline{x}_K \le \eta\cdot\frac{8\epsilon}{\mathcal{R}}\cdot K$ for $K \le \frac{\mathcal{R}^2}{64\epsilon\eta}$. Hence, we can show that $\min\{f(x_K), f(\overline{x}_K)\} - f^* \ge 2\epsilon$ for all K satisfying $K \le \frac{\mathcal{R}^2}{64\epsilon\eta}$. Hence, to reach the desired precision, at least the following number of iterations is necessary:

$$K \ge \tfrac{\mathcal{R}^2}{64\epsilon\eta} \overset{(a)}{\ge} \exp(\tfrac{1}{4}\mathcal{R}\mathcal{G}_1)\cdot\tfrac{\mathcal{R}\mathcal{G}_0}{128\epsilon},$$

where (a) uses the upper bound on η. This concludes Case II. □

H.2 PROOF OF THEOREM 9

This proof is similar in many ways to the proof of Theorem 8 in Section H.1, but is more involved because of the more complex AdaGrad-Norm stepsizes. We consider one-dimensional example

functions $f(x)$. We choose deterministic gradient oracle $\nabla_\xi f(x) = \nabla f(x)$, where $\nabla f(x) \in \partial f(x)$ is an arbitrary subgradient. We further consider the following three cases:

$$\text{Case I: } \eta \geq \mathcal{R}\exp(\tfrac{1}{32}\mathcal{R}\mathcal{G}_1), \quad \text{Case II: } \tfrac{1}{4}\mathcal{R} \leq \eta < \mathcal{R}\exp(\tfrac{1}{32}\mathcal{R}\mathcal{G}_1), \quad \text{Case III: } \eta < \tfrac{1}{4}\mathcal{R}.$$

Case I. We consider the example from Case I in the proof of Theorem 8 in Section H.1. One can verify that $x_K, \overline{x}_K$ coincide with the definitions obtained in Section H.1, Case I. Indeed, we can obtain the following inequality:

$$\|\eta_{K-1}\nabla f(x_{K-1})\|_2 = \eta/\sqrt{K} \geq 2\mathcal{R},$$

as long as K satisfies the following inequality:

$$K \leq \left[\tfrac{\eta}{2\mathcal{R}}\right]^2.$$

Consequently, to reach the precision $\min\{f(x_K), f(\overline{x}_K)\} - f^* \leq \epsilon$, the number of iterations K must satisfy the following inequality:

$$K > \left[\tfrac{\eta}{2\mathcal{R}}\right]^2 \overset{\text{(a)}}{\geq} \tfrac{1}{4}\exp(\tfrac{1}{16}\mathcal{R}\mathcal{G}_1),$$

where (a) uses the lower bound on η. This concludes Case I.

Case II. Let $r = \max\left\{0, \frac{1}{\mathcal{G}_1}\ln\left[\frac{32\epsilon}{\mathcal{R}\mathcal{G}_0}\right]\right\}$, and consider the following example:

$$f(x) = \begin{cases} \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1|x - \frac{1}{16}\mathcal{R}|) - 1] & x \leq \frac{1}{16}\mathcal{R} + r \\ \frac{32\epsilon}{\mathcal{R}}[x - \frac{1}{16}\mathcal{R} - r] + \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] & x > \frac{1}{16}\mathcal{R} + r \end{cases}.$$

Similar to the proof of Case II in Section H.1, one can verify that $f(x)$ is convex and $\nabla f(x)$ satisfies Assumption 2. Additionally, from the assumptions in eq. (34), it follows that $r \leq \frac{1}{32}\mathcal{R}$. Hence, the following inequality holds:

$$f(\tfrac{1}{8}\mathcal{R}) \geq \tfrac{32\epsilon}{\mathcal{R}}[\tfrac{1}{16}\mathcal{R} - r] + \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] = \epsilon + \tfrac{32\epsilon}{\mathcal{R}}[\tfrac{1}{32}\mathcal{R} - r] + \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] > \epsilon.$$

In addition, the following inequality holds:

$$f(0) = \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\tfrac{1}{16}\mathcal{R}\mathcal{G}_1) - 1] \overset{\text{(a)}}{\geq} \tfrac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\tfrac{1}{32}\mathcal{R}\mathcal{G}_1 + 1) - 1] \overset{\text{(b)}}{>} \tfrac{\mathcal{G}_0}{\mathcal{G}_1}\exp(\tfrac{1}{32}\mathcal{R}\mathcal{G}_1) \overset{\text{(c)}}{\geq} \epsilon,$$

where (a), (b) and (c) use the assumptions in eq. (34). Furthermore, from the lower bound on η it follows that $x_1 \geq \frac{1}{4}\mathcal{R}$, and from the upper bound on η it follows that

$$x_K \geq \tfrac{1}{4}\mathcal{R} - \tfrac{\eta}{\|\nabla f(0)\|_2} \cdot \tfrac{32\epsilon}{\mathcal{R}} \cdot (K-1) \geq \tfrac{1}{4}\mathcal{R} - \tfrac{32\epsilon}{\mathcal{G}_0\exp(\frac{1}{32}\mathcal{R}\mathcal{G}_1)} \cdot (K-1) \geq \tfrac{1}{8}\mathcal{R},$$

and $\overline{x}_K \geq \frac{1}{8}\mathcal{R}$ for $K \in \mathbb{N}$, as long as K satisfies the following inequality:

$$K \leq 1 + \exp(\tfrac{1}{32}\mathcal{R}\mathcal{G}_1) \cdot \tfrac{\mathcal{R}\mathcal{G}_0}{256\epsilon}.$$

Hence, to reach the precision $\min\{f(x_K), f(\overline{x}_K)\} - f^* \leq \epsilon$, it is necessary that

$$K \geq \exp(\tfrac{1}{32}\mathcal{R}\mathcal{G}_1) \cdot \tfrac{\mathcal{R}\mathcal{G}_0}{256\epsilon} \overset{\text{(a)}}{\geq} \tfrac{1}{8}\exp(\tfrac{1}{64}\mathcal{R}\mathcal{G}_1),$$

where (a) uses the assumptions in eq. (34). This concludes Case II.

Case III. Let $r = \max\left\{0, \frac{1}{\mathcal{G}_1}\ln\left[\frac{32\epsilon}{\mathcal{R}\mathcal{G}_0}\right]\right\}$, $m = \max\{\mathcal{G}_0, \frac{32\epsilon}{\mathcal{R}}\}$, and consider the following example:

$$f(x) = \begin{cases} 0 & x \geq \mathcal{R} \\ \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1[\mathcal{R} - x]) - 1] & x \in [\mathcal{R} - r, \mathcal{R}) \\ \frac{32\epsilon}{\mathcal{R}}[\mathcal{R} - r - x] + \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] & x \in [\frac{1}{4}\mathcal{R}, \mathcal{R} - r) \\ \frac{m}{\mathcal{G}_1}[\exp(\mathcal{G}_1[\frac{1}{4}\mathcal{R} - x]) - 1] + \frac{32\epsilon}{\mathcal{R}}[\frac{3}{4}\mathcal{R} - r] + \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] & x < \frac{1}{4}\mathcal{R} \end{cases}.$$

One can verify that $f(x)$ is convex, $\nabla f(x)$ satisfies Assumption 2, and $r \leq \frac{1}{32}\mathcal{R}$. Moreover, from the upper bound on η it follows that there exists $\hat{K} \in \mathbb{N}$ such that $x_{\hat{K}} \in (\frac{1}{4}\mathcal{R}, \frac{1}{2}\mathcal{R}]$, and $\overline{x}_K \leq x_{\hat{K}}$ for all $K \leq \hat{K}$. Moreover, we can upper-bound x_K as follows:

$$\begin{aligned} x_K &\leq \tfrac{1}{2}\mathcal{R} + \frac{\eta}{\|\nabla f(0)\|_2} \cdot \frac{32\epsilon}{\mathcal{R}} \cdot (K - \hat{K}) \\ &\leq \tfrac{1}{2}\mathcal{R} + \frac{\mathcal{R}}{4m \exp(\frac{1}{4}\mathcal{R}\mathcal{G}_1)} \cdot \frac{32\epsilon}{\mathcal{R}} \cdot (K - \hat{K}) \\ &\leq \tfrac{1}{2}\mathcal{R} + \frac{\mathcal{R}}{4\mathcal{G}_0 \exp(\frac{1}{4}\mathcal{R}\mathcal{G}_1)} \cdot \frac{32\epsilon}{\mathcal{R}} \cdot (K - \hat{K}) \\ &= \tfrac{1}{2}\mathcal{R} + \frac{8\epsilon}{\mathcal{G}_0 \exp(\frac{1}{4}\mathcal{R}\mathcal{G}_1)} \cdot (K - \hat{K}) \leq \tfrac{3}{4}\mathcal{R}, \end{aligned}$$

as long as K satisfies the following inequality:

$$K \leq \hat{K} + \exp(\tfrac{1}{4}\mathcal{R}\mathcal{G}_1) \cdot \frac{\mathcal{R}\mathcal{G}_0}{32\epsilon},$$

Consequently, for such $K \in \mathbb{N}$, we have

$$\min\{f(x_K), f(\overline{x}_K)\} - f^* \geq f(\tfrac{3}{4}\mathcal{R}) - f^* \geq \frac{32\epsilon}{\mathcal{R}}[\mathcal{R} - r - \tfrac{3}{4}\mathcal{R}] + \frac{\mathcal{G}_0}{\mathcal{G}_1}[\exp(\mathcal{G}_1 r) - 1] \geq 4\epsilon.$$

Hence, to reach the precision $\min\{f(x_K), f(\overline{x}_K)\} - f^* \leq \epsilon$, at least the following number of iterations is necessary:

$$K \geq \hat{K} + \exp(\tfrac{1}{4}\mathcal{R}\mathcal{G}_1) \cdot \frac{\mathcal{R}\mathcal{G}_0}{32\epsilon} \overset{\text{(a)}}{\geq} \exp(\tfrac{1}{8}\mathcal{R}\mathcal{G}_1),$$

where (a) uses the assumptions in eq. (34). This concludes Case III. □

I PROOFS FOR SECTION B

I.1 PROOF OF LEMMA 6

In the case $\beta > 0$, we can upper-bound $\|\nabla h(x)\|_2$ as follows:

$$\begin{aligned}
\|\nabla h(x)\|_2 &\overset{(a)}{=} \tfrac{1}{(1+\nu)[h(x)]^\nu}\|\nabla f(x)\|_2 \\
&\overset{(b)}{\leq} \tfrac{1}{(1+\nu)[h(x)]^\nu}\mathbb{E}_{\xi\sim\mathcal{D}}[\|u_\xi(x)\|_2 + \|v_\xi(x)\|_2] \\
&\overset{(c)}{\leq} \tfrac{1}{(1+\nu)[h(x)]^\nu}\left[\sqrt{\sigma^2 + \mathcal{L}_0^{\frac{2}{1+\nu}}[f(x)-f^*]^{\frac{2\nu}{1+\nu}}} + \mathcal{L}_1[f(x)-f(x^*)]\right] \\
&\overset{(d)}{=} \tfrac{1}{(1+\nu)[h(x)]^\nu}\left[\sqrt{\sigma^2 + \mathcal{L}_0^{\frac{2}{1+\nu}}[h(x)^{1+\nu}-\beta]^{\frac{2\nu}{1+\nu}}} + \mathcal{L}_1[h(x)^{1+\nu}-\beta]\right] \\
&\leq \tfrac{1}{(1+\nu)[h(x)]^\nu}\left[\sigma + \mathcal{L}_0^{\frac{1}{1+\nu}}[h(x)]^\nu - \mathcal{L}_1\beta\right] + \tfrac{1}{1+\nu}\mathcal{L}_1 h(x) \\
&\overset{(e)}{\leq} \tfrac{1}{(1+\nu)[h(x)]^\nu}\left[\max\left\{\mathcal{L}_1\beta, \mathcal{L}_0^{\frac{1}{1+\nu}}\beta^{\frac{\nu}{1+\nu}}\right\} + \mathcal{L}_0^{\frac{1}{1+\nu}}[h(x)]^\nu - \mathcal{L}_1\beta\right] + \tfrac{1}{1+\nu}\mathcal{L}_1 h(x) \\
&\overset{(f)}{\leq} \tfrac{1}{(1+\nu)[h(x)]^\nu}\left[\max\left\{\mathcal{L}_1\beta, \mathcal{L}_0^{\frac{1}{1+\nu}}[h(x)]^\nu\right\} + \mathcal{L}_0^{\frac{1}{1+\nu}}[h(x)]^\nu - \mathcal{L}_1\beta\right] + \tfrac{1}{1+\nu}\mathcal{L}_1 h(x) \\
&\leq \tfrac{2}{(1+\nu)}\mathcal{L}_0^{\frac{1}{1+\nu}} + \tfrac{1}{1+\nu}\mathcal{L}_1 h(x),
\end{aligned}$$

where (a), (d), (e) and (f) use the definitions in eq. (38); (b) uses Assumption 4, Jensen's inequality, and the triangle inequality; (c) uses Assumption 4 and Jensen's inequality. Next, we can upper-bound $|h(x') - h(x)|^{1+\nu}$ as follows:

$$\begin{aligned}
|h(x') - h(x)|^{1+\nu} &\overset{(a)}{\leq} \mathcal{O}\Big[(\mathcal{L}_0^{1/(1+\nu)} + \mathcal{L}_1 h(x))^{1+\nu}\exp(\mathcal{L}_1\|x'-x\|_2)\|x'-x\|_2^{1+\nu}\Big] \\
&\leq \mathcal{O}\Big[\big(\mathcal{L}_0 + [\mathcal{L}_1 h(x)]^{1+\nu}\big)\exp(\mathcal{L}_1\|x'-x\|_2)\|x'-x\|_2^{1+\nu}\Big],
\end{aligned}$$

where (a) uses the inequality above and similar arguments to the proof of Lemma 1 in Section E.1. Finally, the case $\beta = 0$ can be obtained by standard limiting arguments. □

I.2 PROOF OF LEMMA 7

We can upper-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows:

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_K] &\overset{(a)}{\leq} \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \|\nabla_{\xi_k} f_k(z_k)\|_2^2}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2\|\nabla_{\xi_k} f(x_k)\|_2^2}\right] \\
&\overset{(c)}{\leq} \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2\|u_{\xi_k}(x_k) + v_{\xi_k}(x_k)\|_2^2}\right] \\
&\leq \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2\|u_{\xi_k}(x_k)\|_2^2} + \mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2\|v_{\xi_k}(x_k)\|_2^2}\right] \\
&\overset{(d)}{\leq} \mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2\mathbb{E}[\|u_{\xi_k}(x_k)\|_2^2]} + \mathbb{E}\left[\mathcal{C}\mathcal{R}\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2\|v_{\xi_k}(x_k)\|_2^2}\right] \\
&\overset{(e)}{\leq} \mathcal{C}\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \mathcal{C}\mathcal{R}\mathcal{L}_0^{\frac{1}{1+\nu}}\sqrt{\mathbb{E}\left[\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2[f(x_k)-f^*]^{\frac{2\nu}{1+\nu}}\right]} \\
&\quad + \mathbb{E}\left[\mathcal{C}\mathcal{R}\mathcal{L}_1\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2[f(x_k)-f^*]^2}\right] \\
&\overset{(f)}{\leq} \mathcal{C}\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \mathcal{C}\mathcal{R}\mathcal{L}_0^{\frac{1}{1+\nu}}\sqrt{\textstyle\sum_{k=1}^K [\alpha_k\pi_k]^{\frac{2}{1+\nu}}[\alpha_k\pi_k\mathbb{E}[f(x_k)-f^*]]^{\frac{2\nu}{1+\nu}}}
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\mathcal{C}\mathcal{R}\mathcal{L}_1\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2[f(x_k)-f^*]^2}\right] \\
&\overset{(g)}{\leq} \mathcal{C}\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \mathbb{E}\left[\mathcal{C}\mathcal{R}\mathcal{L}_1\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2[f(x_k)-f^*]^2}\right] \\
&+ \mathcal{C}\mathcal{R}\mathcal{L}_0^{\frac{1}{1+\nu}}\left[\textstyle\sum_{k=1}^K[\alpha_k\pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2(1+\nu)}}\left[\textstyle\sum_{k=1}^K \alpha_k\pi_k\mathbb{E}[f(x_k)-f^*]\right]^{\frac{\nu}{1+\nu}} \\
&\overset{(h)}{\leq} \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \mathcal{L}_0\mathcal{R}^{1+\nu}\left[\textstyle\sum_{k=1}^K[\alpha_k\pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right] \\
&+ \tfrac{1}{4}\textstyle\sum_{k=1}^K \alpha_k\pi_k\mathbb{E}[f(x_k)-f^*] + \mathcal{O}\left[\mathbb{E}\left[\mathcal{R}\mathcal{L}_1\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2[f(x_k)-f^*]^2}\right]\right],
\end{aligned}
$$

where (a) uses Assumption 3 and the definition in eq. (19); (b) uses eq. (18) and the definition in eq. (17); (c) and (e) use Assumption 4; (d) and (f) use Jensen's inequality; (g) uses Hölder's inequality; (h) uses Young's inequality.

Next, we define $J \in \mathbb{N}$, $k_s \in \mathbb{N}$ and $\mathcal{K}_s \subset \mathbb{N}$ for $s \in \{1, \ldots, J\}$ similarly to the proof of Lemma 4 in Section G.5. Consequently, it holds that $\pi_k \leq \mathcal{O}[\pi_{k_s}]$. Furthermore, for $k \in \mathcal{K}_s$, using the fact that $\alpha_k \leq \alpha_{k_s}$ due to the definition in eq. (41), we can upper-bound $\|x_k - x_{k_s}\|_2$ as follows:

$$\|x_k - x_{k_s}\|_2 \leq \mathcal{O}[\alpha_{k_s} T\mathcal{R}] \leq \mathcal{O}[1/\mathcal{L}_1].$$

Furthermore, using similar arguments to the proof of Lemma 5 in Section G.7, we get the following:

$$\textstyle\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^q \leq \mathcal{O}\big[\sum_{k\in\mathcal{K}_s} \alpha_k^q\big] \quad \text{for all} \;\; q > 0 \text{ and } s \in \{1, \ldots, J\}.$$

Next, we can obtain the following inequality:

$$
\begin{aligned}
&\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2[f(x_k)-f^*]^2} = \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[f(x_k)-f^*]^2} \\
&\leq \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[f(x_{k_s})-f^*]^2} + \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[f(x_k)-f(x_{k_s})]^2} \\
&\overset{(a)}{\leq} \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[h(x_{k_s})]^{2(1+\nu)}} + \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[[h(x_k)]^{1+\nu}-[h(x_{k_s})]^{1+\nu}]^2} \\
&\leq \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[h(x_{k_s})]^{2(1+\nu)}} \\
&+ \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[h(x_{k_s})+|h(x_k)-h(x_{k_s})|]^{2\nu}[h(x_k)-h(x_{k_s})]^2}\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[h(x_{k_s})]^{2(1+\nu)}} + \sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2|h(x_k)-h(x_{k_s})|^{2(1+\nu)}}\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[h(x_{k_s})]^{2(1+\nu)}}\right] \\
&+ \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2\|x_k-x_{k_s}\|_2^{2(1+\nu)}(\mathcal{L}_0+[\mathcal{L}_1 h(x_{k_s})]^{1+\nu})^2}\right] \\
&\overset{(d)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_k^2\pi_k^2[h(x_{k_s})]^{2(1+\nu)}} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^{2(1+\nu)}\alpha_k^2\pi_{k_s}^2}\right] \\
&\overset{(e)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^2\pi_{k_s}^2[h(x_{k_s})]^{2(1+\nu)}} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^{4+2\nu}\pi_{k_s}^2}\right] \\
&\overset{(f)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^2\pi_{k_s}^2[h(x_{k_s})]^{2(1+\nu)}} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\textstyle\sum_{s=1}^J\sqrt{T}\alpha_{k_s}^{2+\nu}\pi_{k_s}\right] \\
&\overset{(g)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^2\pi_{k_s}^2[h(x_{k_s})]^{2(1+\nu)}} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\tfrac{1}{\sqrt{T}}\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s}\alpha_{k_s}^{2+\nu}\pi_{k_s}\right] \\
&\overset{(h)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^2\pi_{k_s}^2[h(x_{k_s})]^{2(1+\nu)}} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\tfrac{1}{\sqrt{T}}\textstyle\sum_{k=1}^K\alpha_k^{2+\nu}\pi_k\right] \\
&\overset{(i)}{\leq} \mathcal{O}\left[\sqrt{\textstyle\sum_{s=1}^J\sum_{k\in\mathcal{K}_s} \alpha_{k_s}^2\pi_{k_s}^2[f(x_{k_s})-f^*]^2} + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\tfrac{1}{\sqrt{T}}\textstyle\sum_{k=1}^K\alpha_k^{2+\nu}\pi_k + \beta\sqrt{\textstyle\sum_{k=1}^K\alpha_k^2\pi_k^2}\right]
\end{aligned}
$$

$$\overset{(j)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sqrt{T}\alpha_{k_s}\pi_{k_s}[f(x_{k_s})-f^*] + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\frac{1}{\sqrt{T}}\sum_{k=1}^{K}\alpha_k^{2+\nu}\pi_k + [\sigma/\mathcal{L}_1]\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2}\right],$$

where (a) and (i) uses the definition in eq. (38); (b) uses Young's inequality; (c) uses Lemma 6 and the upper-bound on $\|x_k - x_{k_s}\|_2$ above; (d) uses the upper bounds on π_k and $\|x_k - x_{k_s}\|_2$; (e) uses the inequalities $\pi_k \leq \mathcal{O}[\pi_{k_s}]$ and $\alpha_k \leq \alpha_{k_s}$; (f) and (g) use the properties of $\mathcal{K}_s$; (h) uses the inequality above with $q = 2+\nu$ and the fact that $\pi_{k_s} \leq \pi_k$; (j) uses the definition in eq. (38) and the properties of $\mathcal{K}_s$. Furthermore, we can obtain the following inequality:

$$\begin{aligned}
&\sum_{s=1}^{J}\sqrt{T}\alpha_{k_s}\pi_{k_s}[f(x_{k_s})-f^*] \\
&\overset{(a)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[[h(x_{k_s})]^{1+\nu} - [h(x_k)]^{1+\nu} + f(x_k) - f^*\right]\right] \\
&\leq \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[[h(x_k) + |h(x_{k_s}) - h(x_k)|]^{\nu}|h(x_{k_s}) - h(x_k)| + f(x_k) - f^*\right]\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[|h(x_{k_s}) - h(x_k)|^{1+\nu} + \beta^{\frac{\nu}{1+\nu}}|h(x_{k_s}) - h(x_k)| + f(x_k) - f^*\right]\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\|x_k - x_{k_s}\|_2^{1+\nu}\left(\mathcal{L}_0 + [\mathcal{L}_1 h(x_k)]^{1+\nu}\right)\right] \\
&+ \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[\beta^{\frac{\nu}{1+\nu}}\|x_k - x_{k_s}\|_2\left(\mathcal{L}_0^{\frac{1}{1+\nu}} + \mathcal{L}_1 h(x_k)\right) + f(x_k) - f^*\right]\right] \\
&\overset{(d)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\|x_k - x_{k_s}\|_2^{1+\nu}\left(\mathcal{L}_0 + \mathcal{L}_1^{1+\nu}\beta + \mathcal{L}_1^{1+\nu}[f(x_k) - f^*]\right)\right] \\
&+ \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[\|x_k - x_{k_s}\|_2\left([\mathcal{L}_0\beta^{\nu}]^{\frac{1}{1+\nu}} + \mathcal{L}_1\beta + \mathcal{L}_1[f(x_k) - f^*]\right) + f(x_k) - f^*\right]\right] \\
&\overset{(e)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\|x_k - x_{k_s}\|_2^{1+\nu}\left(\mathcal{L}_0 + \sigma\mathcal{L}_1^{\nu} + \mathcal{L}_1^{1+\nu}[f(x_k) - f^*]\right)\right] \\
&+ \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}[\|x_k - x_{k_s}\|_2(\sigma + \mathcal{L}_1[f(x_k) - f^*]) + f(x_k) - f^*]\right] \\
&\overset{(f)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[\|x_k - x_{k_s}\|_2^{1+\nu}(\mathcal{L}_0 + \sigma\mathcal{L}_1^{\nu}) + \|x_k - x_{k_s}\|_2\sigma + f(x_k) - f^*\right]\right] \\
&\overset{(g)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}\left[[\alpha_{k_s}T\mathcal{R}]^{1+\nu}(\mathcal{L}_0 + \sigma\mathcal{L}_1^{\nu}) + \alpha_{k_s}T\mathcal{R}\sigma + f(x_k) - f^*\right]\right] \\
&\overset{(h)}{\leq} \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\pi_{k_s}}{\sqrt{T}}\left[\left[\alpha_k^{2+\nu} + \alpha_{k_s}^{2+\nu}\right][T\mathcal{R}]^{1+\nu}(\mathcal{L}_0 + \sigma\mathcal{L}_1^{\nu}) + [\alpha_k^2 + \alpha_{k_s}^2]T\mathcal{R}\sigma\right]\right] \\
&+ \mathcal{O}\left[\sum_{s=1}^{J}\sum_{k\in\mathcal{K}_s}\frac{\alpha_k\pi_{k_s}}{\sqrt{T}}[f(x_k) - f^*]\right] \\
&\overset{(i)}{\leq} \mathcal{O}\left[\sum_{k=1}^{K}\frac{\alpha_k\pi_k}{\sqrt{T}}\left[[\alpha_k T\mathcal{R}]^{1+\nu}(\mathcal{L}_0 + \sigma\mathcal{L}_1^{\nu}) + \alpha_k T\mathcal{R}\sigma + f(x_k) - f^*\right]\right],
\end{aligned}$$

where (a) uses the definition in eq. (38) and the properties of $\mathcal{K}_s$ and the inequality above with $q = 1$; (b) and (d) use the definitions in eq. (38) and Young's inequality; (c) uses Lemma 6 and the upper bound on $\|x_k - x_{k_s}\|_2$ above; (e) uses the definition in eq. (38); (f) and (g) use the upper bound on $\|x_k - x_{k_s}\|_2$ above; (h) uses Young's inequality; (i) uses the inequality above with $q = 2+\nu$ and $q = 2$, and the fact that $\pi_{k_s} \leq \pi_k$. Finally, we can upper-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows:

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_K] &\overset{(a)}{\leq} \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathcal{L}_0\mathcal{R}^{1+\nu}\left[\sum_{k=1}^{K}[\alpha_k\pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right] \\
&\quad + \tfrac{1}{4}\sum_{k=1}^{K}\alpha_k\pi_k\mathbb{E}[f(x_k) - f^*] + \mathcal{O}\left[\mathbb{E}\left[\mathcal{R}\mathcal{L}_1\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2[f(x_k) - f^*]^2}\right]\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathcal{L}_0\mathcal{R}^{1+\nu}\left[\sum_{k=1}^{K}[\alpha_k\pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right] + \tfrac{1}{4}\sum_{k=1}^{K}\alpha_k\pi_k\mathbb{E}[f(x_k) - f^*] \\
&\quad + \mathcal{O}\left[\mathcal{R}\mathcal{L}_1\sum_{s=1}^{J}\sqrt{T}\alpha_{k_s}\pi_{k_s}\mathbb{E}[f(x_{k_s}) - f^*] + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0\frac{\mathcal{R}\mathcal{L}_1}{\sqrt{T}}\sum_{k=1}^{K}\alpha_k^{2+\nu}\pi_k\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\sum_{k=1}^{K}\alpha_k^2\pi_k^2} + \mathcal{L}_0\mathcal{R}^{1+\nu}\left[\sum_{k=1}^{K}[\alpha_k\pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathcal{O}\left[\mathcal{R}\mathcal{L}_1 \textstyle\sum_{k=1}^K \frac{\alpha_k \pi_k}{\sqrt{T}} \left[[\alpha_k T\mathcal{R}]^{1+\nu}(\mathcal{L}_0 + \sigma\mathcal{L}_1^\nu) + \alpha_k T\mathcal{R}\sigma\right]\right] \\
&+ \mathcal{O}\left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0 \frac{\mathcal{R}\mathcal{L}_1}{\sqrt{T}} \textstyle\sum_{k=1}^K \alpha_k^{2+\nu}\pi_k\right] + \left[\frac{1}{4} + \mathcal{O}\left[\frac{\mathcal{R}\mathcal{L}_1}{\sqrt{T}}\right]\right] \textstyle\sum_{k=1}^K \alpha_k \pi_k \mathbb{E}[f(x_k) - f^*] \\
&\overset{(d)}{\leq} \tfrac{1}{2}\textstyle\sum_{k=1}^K \alpha_k \pi_k \mathbb{E}[f(x_k) - f^*] + \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathcal{L}_0 \mathcal{R}^{1+\nu}\left[\textstyle\sum_{k=1}^K [\alpha_k \pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right] \\
&+ \mathcal{O}\left[T\mathcal{R}\sigma \textstyle\sum_{k=1}^K \alpha_k^2 \pi_k ([\alpha_k T\mathcal{R}\mathcal{L}_1]^\nu + 1) + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0 \textstyle\sum_{k=1}^K \alpha_k^{2+\nu}\pi_k\right] \\
&\overset{(e)}{\leq} \tfrac{1}{2}\textstyle\sum_{k=1}^K \alpha_k \pi_k \mathbb{E}[f(x_k) - f^*] + \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathcal{L}_0 \mathcal{R}^{1+\nu}\left[\textstyle\sum_{k=1}^K [\alpha_k \pi_k]^{\frac{2}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right] \\
&+ \mathcal{O}\left[T\mathcal{R}\sigma \textstyle\sum_{k=1}^K \alpha_k^2 \pi_k + [T\mathcal{R}]^{1+\nu}\mathcal{L}_0 \textstyle\sum_{k=1}^K \alpha_k^{2+\nu}\pi_k\right],
\end{aligned}
$$

where (a), (b) and (c) use the inequalities above; (d) uses the definition of T; (e) uses the definition in eq. (41). □

I.3 Proof of Theorem 10

First, observe that the statement of Lemma 7 holds for arbitrary $K \geq T$. Next, we define $\Delta_k = \mathbb{E}[f(x_k) - f^*]$ and δ_k as follows:

$$
\begin{aligned}
\delta_k = e\mathcal{F} &+ \mathcal{O}\left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0 \textstyle\sum_{i=1}^k \alpha_i^{2+\nu}\pi_i + T\mathcal{R}\sigma\textstyle\sum_{i=1}^k \alpha_i^2 \pi_i\right] \\
&+ \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\textstyle\sum_{i=1}^k \alpha_i^2 \pi_i^2} + \mathcal{L}_0 \mathcal{R}^{1+\nu}\left[\textstyle\sum_{i=1}^k [\alpha_i \pi_i]^{\frac{2}{1-\nu}}\right]^{(1-\nu)/2}\right].
\end{aligned}
$$

Using Lemmas 2 and 7, and using similar arguments to Section G.6, we can show that the following inequality holds for all $k \in \mathbb{N}$:

$$
\pi_k \Delta_k \leq \delta_k + \tfrac{1}{2}\textstyle\sum_{i=1}^k \alpha_i \pi_i \Delta_i.
$$

Hence, the conditions of Lemma 10 with $p = 1/2$ are satisfied. Furthermore, for $k \leq ST$, we can upper-bound δ_k as follows:

$$
\begin{aligned}
\delta_k &\overset{(a)}{\leq} e\mathcal{F} + \mathcal{O}\left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0 \alpha_1^{2+\nu}\pi_k \textstyle\sum_{i=0}^\infty (1-\alpha_1)^i + T\mathcal{R}\sigma\alpha_1^2\pi_k\textstyle\sum_{i=0}^\infty (1-\alpha_1)^i\right] \\
&\quad + \mathcal{O}\left[\mathcal{R}\sigma\alpha_1\pi_k\sqrt{\textstyle\sum_{i=0}^\infty (1-\alpha_1)^{2i}} + \mathcal{L}_0\mathcal{R}^{1+\nu}\alpha_1\pi_k\left[\textstyle\sum_{i=0}^\infty (1-\alpha_1)^{\frac{2i}{1-\nu}}\right]^{(1-\nu)/2}\right] \\
&\leq e\mathcal{F} + \mathcal{O}\left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0\alpha_1^{1+\nu}\pi_k + T\mathcal{R}\sigma\alpha_1\pi_k\right] \\
&\quad + \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\alpha_1}\pi_k + \mathcal{L}_0\mathcal{R}^{1+\nu}\alpha_1^{\frac{1+\nu}{2}}\pi_k\right] \\
&\overset{(b)}{\leq} e\mathcal{F} + \mathcal{O}\left[[\mathcal{L}_0/\mathcal{L}_1^{1+\nu}]\pi_k + [\sigma/\mathcal{L}_1]\pi_k\right] \\
&\overset{(c)}{=} e\mathcal{F} + \mathcal{O}[\hat{\mathcal{F}}\pi_k]
\end{aligned}
$$

where (a) uses the definitions in eqs. (17) and (41); (b) uses the definition in eq. (41); (c) uses the definition of $\hat{\mathcal{F}}$. For $k \geq ST + 1$, we can upper-bound δ_k as follows:

$$
\begin{aligned}
\delta_k &\overset{(a)}{\leq} \delta_{ST} + \mathcal{O}\left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0\alpha_{ST+1}^{2+\nu}\pi_k\textstyle\sum_{i=0}^\infty (1-\alpha_{ST+1})^i + T\mathcal{R}\sigma\alpha_{ST+1}^2\pi_k\textstyle\sum_{i=0}^\infty (1-\alpha_{ST+1})^i\right] \\
&+ \mathcal{O}\left[\mathcal{R}\sigma\alpha_{ST+1}\pi_k\sqrt{\textstyle\sum_{i=0}^\infty (1-\alpha_{ST+1})^{2i}} + \mathcal{L}_0\mathcal{R}^{1+\nu}\alpha_{ST+1}\pi_k\left[\textstyle\sum_{i=0}^\infty (1-\alpha_{ST+1})^{\frac{2i}{1-\nu}}\right]^{\frac{1-\nu}{2}}\right] \\
&\leq \delta_{ST} + \mathcal{O}\left[[T\mathcal{R}]^{1+\nu}\mathcal{L}_0\alpha_{ST+1}^{1+\nu}\pi_k + T\mathcal{R}\sigma\alpha_{ST+1}\pi_k\right] \\
&+ \mathcal{O}\left[\mathcal{R}\sigma\sqrt{\alpha_{ST+1}}\pi_k + \mathcal{L}_0\mathcal{R}^{1+\nu}\alpha_{ST+1}^{\frac{1+\nu}{2}}\pi_k\right] \\
&\overset{(b)}{\leq} \delta_{ST} + \mathcal{O}[\epsilon\pi_k],
\end{aligned}
$$

where (a) uses the definitions in eqs. (17) and (41); (b) uses the definition in eq. (41). Finally, we can upper-bound Δ_K as follows:

$$\begin{aligned}
\Delta_K &\overset{(a)}{\leq} \mathcal{O}\left[\frac{\delta_K}{\pi_K} + \sum_{k=1}^{K} \frac{\alpha_k \delta_k}{\sqrt{\pi_K \pi_{k-1}}}\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\frac{\delta_K}{\pi_K} + \sum_{k=1}^{ST} \frac{\alpha_k \delta_k}{\sqrt{\pi_K \pi_{k-1}}} + \sum_{k=ST+1}^{K} \frac{\alpha_k \delta_k}{\sqrt{\pi_K \pi_{k-1}}}\right] \\
&\overset{(c)}{\leq} \mathcal{O}\left[\frac{\mathcal{F}}{\pi_K} + \frac{\hat{\mathcal{F}} \pi_{ST}}{\pi_K} + \epsilon + \sum_{k=1}^{ST}\left[\frac{\alpha_k \mathcal{F}}{\sqrt{\pi_K \pi_{k-1}}} + \frac{\hat{\mathcal{F}} \alpha_k \pi_k}{\sqrt{\pi_K \pi_{k-1}}}\right]\right] \\
&\quad + \mathcal{O}\left[\sum_{k=ST+1}^{K}\left[\frac{\alpha_k \mathcal{F}}{\sqrt{\pi_K \pi_{k-1}}} + \frac{\hat{\mathcal{F}} \alpha_k \pi_{ST}}{\sqrt{\pi_K \pi_{k-1}}} + \frac{\epsilon \alpha_k \pi_k}{\sqrt{\pi_K \pi_{k-1}}}\right]\right] \\
&\overset{(d)}{\leq} \mathcal{O}\left[\mathcal{F}[1-\alpha_1]^{ST/2}[1-\alpha_{ST+1}]^{(K-ST)/2} + \hat{\mathcal{F}}[1-\alpha_{ST+1}]^{(K-ST)/2} + \epsilon\right] \\
&\overset{(e)}{\leq} \mathcal{O}[\epsilon],
\end{aligned}$$

where (a) uses Lemma 10 with $p = 1/2$; (b) uses the fact that $K \geq TS + 1$; (c) uses the inequalities above; (d) uses the definitions in eqs. (17) and (41); (e) uses the definitions in eq. (42). □

J Proofs for Section C

J.1 Technical Lemma

In this section, we present a technical lemma that will be used in the proof of Theorem 11.

Lemma 11. *Let $p \in (0, 1/2]$, and let $\{\Delta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ and $\{\delta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ be two sequences of non-negative numbers satisfying the following condition for all $k \in \mathbb{N}$:*

$$\pi_k\Delta_k \leq \delta_k + p\gamma\textstyle\sum_{i=1}^k \alpha_i\pi_i\Delta_i, \tag{58}$$

where α_k, π_k satisfy eq. (51). Then, the following inequalities hold for all $k \in \mathbb{N}$:

$$\textstyle\sum_{i=1}^k \alpha_i\pi_i\Delta_i \leq \sum_{i=1}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p. \tag{59}$$

Proof. We define $\tau_k \geq 0$ for $k \in \mathbb{N}$ as follows:

$$\tau_k = \textstyle\sum_{i=1}^k \alpha_i\pi_i\Delta_i.$$

We can upper-bound τ_k as follows:

$$\tau_k \overset{\text{(a)}}{=} \alpha_k\pi_k\Delta_k + \tau_{k-1} \overset{\text{(b)}}{\leq} \alpha_k\delta_k + p\gamma\alpha_k\tau_k + \tau_{k-1} \overset{\text{(c)}}{\leq} \alpha_k\delta_k + [1-(1-\gamma\alpha_k)^p]\tau_k + \tau_{k-1},$$

where (a) uses the definition of τ_k; (b) uses the assumption of Lemma 11; (c) uses the inequality $pt \leq 1-(1-t)^p$ for $t < 1$ which is implied by the concavity of the function $t \mapsto t^p$. After rearranging, we obtain the following:

$$\alpha_k\delta_k + \tau_{k-1} \geq (1-\gamma\alpha_k)^p\tau_k \overset{\text{(a)}}{=} [\pi_{k-1}/\pi_k]^p\tau_k$$

where (a) uses the definition in eq. (51). After rearranging one more time, we obtain the following:

$$[1/\pi_k]^p\tau_k \leq [1/\pi_{k-1}]^p[\tau_{k-1} + \alpha_k\delta_k].$$

Hence, we can upper-bound τ_k as follows:

$$\tau_k \leq \textstyle\sum_{i=1}^k \alpha_i\delta_i[\pi_k/\pi_{i-1}]^p.$$

□

J.2 Proof of Lemma 8

Equation (44) implies Equation (48). Since the function $f^\circ(x;\cdot)$ is subadditive and homogeneous, it is equal to the support function of the generalized subdifferential $\partial^\circ f(x)$. Hence, $\langle w, g\rangle \leq f^\circ(x; w)$ for all $g \in \partial^\circ f(x)$. Furthermore, we can upper-bound $\langle x^* - x, \nabla f(x; w)\rangle$ as follows:

$$\begin{aligned}
\langle x^* - x, \nabla f(x; w)\rangle &\overset{\text{(a)}}{\leq} f^\circ(x; x^* - x) \\
&\overset{\text{(b)}}{=} \limsup_{x'\to x, t\to+0} \tfrac{1}{t}[f(x' + t(x^* - x)) - f(x')] \\
&\leq \limsup_{x'\to x, t\to+0} \tfrac{1}{t}[f(x' + t(x^* - x)) - f(x' + t(x^* - x'))] \\
&\quad + \limsup_{x'\to x, t\to+0} \tfrac{1}{t}[f(x' + t(x^* - x')) - f(x')] \\
&\overset{\text{(c)}}{=} \limsup_{x'\to x, t\to+0} \tfrac{1}{t}[f(x' + t(x^* - x')) - f(x')] \\
&\overset{\text{(d)}}{\leq} \limsup_{x'\to x, t\to+0} \gamma[f^* - f(x')] \\
&\overset{\text{(e)}}{=} \gamma[f^* - f(x)],
\end{aligned}$$

where (a) uses the fact that $\nabla f(x; w) \in \partial^\circ f(x)$; (b) uses the definition in eq. (45); (c) and (e) use the fact that the function $f(x)$ is locally Lipschitz; (d) uses the γ-quasar convexity of the function $f(x)$ in eq. (44).

Equation (48) implies Equation (44). Let the function $p(t)\colon [0, 1] \to \mathbb{R}_+$ be defined as follows:

$$p(t) = f(x^* + t(x - x^*)) - f^*.$$

We can lower-bound $p(t') - p(t)$ for all $0 < t < t' \leq 1$ as follows:

$$\begin{aligned}
p(t') - p(t) &\overset{(a)}{=} \int_0^1 f'(x^* + (t + s(t'-t))(x - x^*); (t'-t)(x - x^*))\mathrm{d}s \\
&\overset{(b)}{=} \int_0^1 \langle \nabla f(x^* + (t + s(t'-t))(x - x^*); (t'-t)(x - x^*)), (t'-t)(x - x^*)\rangle \mathrm{d}s \\
&= \int_0^1 \frac{\langle \nabla f(x^* + (t + s(t'-t))(x - x^*); (t'-t)(x - x^*)), [x^* + (t + s(t'-t))(x - x^*)] - x^*\rangle \mathrm{d}s}{t/(t'-t) + s} \\
&\overset{(c)}{\geq} \int_0^1 \frac{\gamma[f(x^* + (t + s(t'-t))(x - x^*)) - f^*](t'-t)\mathrm{d}s}{t + s(t'-t)} = \int_t^{t'} \frac{\gamma}{s} p(s)\mathrm{d}s \geq 0,
\end{aligned}$$

where (a) uses Lemma 3 of Zhang et al. (2020); (b) uses the property in eq. (47); (c) uses the inequality in eq. (48). Hence, the function $p(t)$ is non-decreasing. Furthermore, let $c \in [0, 1]$ such that $p(c) > 0$, and let the function $q(t) \colon [c, 1] \to \mathbb{R}$ be defined as follows:

$$q(t) = p(c) + \int_c^t \frac{\gamma}{s} p(s) ds.$$

From the previously obtained inequalities, we can conclude that $p(c) = q(c) \leq q(t) \leq p(t)$ for all $t \in [c, 1]$. Moreover, from the continuity of the function $p(t)$ it follows that the function $q(t)$ is continuously differentiable, and we can lower-bound its derivative $\dot{q}(t)$ as follows:

$$\dot{q}(t) \overset{(a)}{=} [\gamma/t] \cdot p(t) \overset{(b)}{\geq} [\gamma/t] \cdot q(t),$$

where (a) uses the definition of the function $q(t)$; (b) uses the inequality $q(t) \leq p(t)$ above. This implies the following inequality:

$$\dot{l}(t) \geq \gamma/t,$$

where $l(t) = \ln q(t)$. After integrating this inequality on $[c, 1]$, we obtain the following:

$$\ln q(1) \geq \ln q(c) - \gamma \ln c.$$

By taking the exponent of both sides and rearranging, we obtain the following:

$$q(c) \leq c^\gamma q(1) \leq (1 - \gamma(1 - c))q(1).$$

Finally, we can upper-bound $p(c)$ as follows:

$$p(c) = q(c) \leq (1 - \gamma(1 - c))q(1) \leq (1 - \gamma(1 - c))p(1),$$

which trivially holds even if $p(c) = 0$ and proves the inequality in eq. (44) by choosing $c = 1 - t$. □

J.3 Proof of Theorem 11

First, using the definitions in eqs. (18), (50) and (52) and the fact that $z_k \in Q_\mathcal{R}$, we can upper-bound $\|\overline{x}_k - x_k\|_2$ and $\|x_k - x_{k-1}\|_2$ as follows:

$$\max\{\|\overline{x}_k - x_k\|_2, \|x_k - x_{k-1}\|_2\} \overset{\text{a.s.}}{\leq} \mathcal{O}[\alpha_k \mathcal{R}] \leq \mathcal{O}[1/(T\mathcal{G}_1)].$$

Next, we can upper-bound $(1 - \alpha_k)\pi_k \mathbb{E}[f(x_k) - f(x_{k-1})]$ as follows:

$$\begin{aligned}
&(1 - \alpha_k)\pi_k \mathbb{E}[f(x_k) - f(x_{k-1})] \\
&\quad \overset{(a)}{=} (1 - \alpha_k)\pi_k \mathbb{E}\left[\int_0^1 \langle \nabla f(x_{k-1} + \zeta(x_k - x_{k-1}); x_k - x_{k-1}), x_k - x_{k-1}\rangle \mathrm{d}\zeta\right] \\
&\quad \overset{(b)}{=} (1 - \alpha_k)\pi_k \mathbb{E}[\langle \nabla f(\overline{x}_k; x_k - x_{k-1}), x_k - x_{k-1}\rangle] \\
&\quad \overset{(c)}{=} \alpha_k \pi_k \mathbb{E}[\langle \nabla f(\overline{x}_k; x_k - x_{k-1}), z_k - x_k\rangle] \\
&\quad = \alpha_k \pi_k \mathbb{E}[\langle \nabla f(\overline{x}_k; x_k - x_{k-1}), x^* - \overline{x}_k + \overline{x}_k - x_k + z_k - x^*\rangle] \\
&\quad \overset{(d)}{\leq} \alpha_k \pi_k \mathbb{E}[\gamma[f^* - f(\overline{x}_k)] + \langle \nabla f(\overline{x}_k; x_k - x_{k-1}), \overline{x}_k - x_k + z_k - x^*\rangle] \\
&\quad \overset{(e)}{=} \gamma \alpha_k \pi_k \mathbb{E}[f^* - f(\overline{x}_k)] + \alpha_k \pi_k \mathbb{E}[\langle \nabla f(\overline{x}_k; x_k - x_{k-1}), \overline{x}_k - x_k\rangle] + \mathbb{E}[\langle g_k, z_k - x^*\rangle],
\end{aligned}$$

where (a) Lemma 3 of Zhang et al. (2020) and the property in eq. (47); (b) uses the definition of $\overline{x}_k$ in eq. (50) and the definition of ζ_k and its independence of x_k and x_{k-1}; (c) uses the definition in eq. (18); (d) uses Lemma 8; (e) uses the definition of g_k in eq. (50), the unbiasedness in Assumption 5, and the fact that ξ_k is independent of ζ_k and z_k, x_k, x_{k-1}. After rearranging, we obtain the following inequality:

$$
\begin{aligned}
&\pi_k \mathbb{E}[f(x_k) - f^*] \\
&\quad \leq (1-\gamma\alpha_k)\pi_k \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] \\
&\quad + (1-\gamma)\alpha_k \pi_k \mathbb{E}[f(x_k) - f(x_{k-1})] + \gamma\alpha_k\pi_k \mathbb{E}[f(x_k) - f(\overline{x}_k)] \\
&\quad + \alpha_k \pi_k \mathbb{E}[\langle \nabla f(\overline{x}_k; x_k - x_{k-1}), \overline{x}_k - x_k \rangle] \\
&\quad \overset{(a)}{\leq} \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] \\
&\quad + (1-\gamma)\alpha_k \pi_k \mathbb{E}[f(x_k) - f(x_{k-1})] + \gamma\alpha_k\pi_k \mathbb{E}[f(x_k) - f(\overline{x}_k)] \\
&\quad + \alpha_k \pi_k \mathbb{E}[\langle \nabla f(\overline{x}_k; x_k - x_{k-1}), \overline{x}_k - x_k \rangle] \\
&\quad \overset{(b)}{\leq} \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] \\
&\quad + (1-\gamma)\alpha_k \pi_k \mathbb{E}[f(x_k) - f(x_{k-1})] + \gamma\alpha_k\pi_k \mathbb{E}[f(x_k) - f(\overline{x}_k)] \\
&\quad + \mathcal{O}\big[\alpha_k^2 \pi_k \mathcal{R}\mathcal{G}_0 + \tfrac{\alpha_k \pi_k}{T} \mathbb{E}[f(\overline{x}_k) - f^*]\big] \\
&\quad \leq \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] + \mathcal{O}\big[\alpha_k^2 \pi_k \mathcal{R}\mathcal{G}_0 + \tfrac{\alpha_k \pi_k}{T} \mathbb{E}[f(x_k) - f^*]\big] \\
&\quad + \mathcal{O}[\alpha_k \pi_k \mathbb{E}[|f(x_k) - f(x_{k-1})| + |f(x_k) - f(\overline{x}_k)|]] \\
&\quad \overset{(c)}{\leq} \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] + \mathcal{O}\big[\alpha_k^2 \pi_k \mathcal{R}\mathcal{G}_0 + \tfrac{\alpha_k \pi_k}{T} \mathbb{E}[f(x_k) - f^*]\big] \\
&\quad + \mathcal{O}[\alpha_k \pi_k \mathbb{E}[(\mathcal{G}_0 + \mathcal{G}_1[f(x_k) - f^*]) \exp(\mathcal{G}_1 \|\overline{x}_k - x_k\|_2) \|\overline{x}_k - x_k\|_2]] \\
&\quad + \mathcal{O}[\alpha_k \pi_k \mathbb{E}[(\mathcal{G}_0 + \mathcal{G}_1[f(x_k) - f^*]) \exp(\mathcal{G}_1 \|x_{k-1} - x_k\|_2) \|x_{k-1} - x_k\|_2]] \\
&\quad \overset{(d)}{\leq} \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] + \mathcal{O}\big[\alpha_k^2 \pi_k \mathcal{R}\mathcal{G}_0 + \tfrac{\alpha_k \pi_k}{T} \mathbb{E}[f(x_k) - f^*]\big] \\
&\quad \overset{(e)}{\leq} \pi_{k-1} \mathbb{E}[f(x_{k-1}) - f^*] + \mathbb{E}[\langle g_k, z_k - x^* \rangle] + \tfrac{1}{4}\gamma\alpha_k\pi_k \mathbb{E}[f(x_k) - f^*] + \mathcal{O}\big[\alpha_k^2 \pi_k \mathcal{R}\mathcal{G}_0\big]
\end{aligned}
$$

where (a) uses the definition in eq. (51); (b) uses Assumption 5 and the upper bound on $\|\overline{x}_k - x_k\|_2$ above; (c) uses Assumption 5, Lemma 9, and Lemma 1; (d) uses the upper bounds on $\|\overline{x}_k - x_k\|_2$ and $\|x_{k-1} - x_k\|_2$ above; (e) uses the definition of T. After summing these inequalities for $k = 1, \ldots, K$, we obtain the following:

$$
\begin{aligned}
\pi_K \mathbb{E}[f(x_K) - f^*] \leq [f(x_0) - f^*] + \mathbb{E}[\mathrm{Reg}_K] \\
+ \tfrac{1}{4}\gamma \textstyle\sum_{k=1}^K \alpha_k \pi_k \mathbb{E}[f(x_k) - f^*] + \mathcal{O}\Big[\mathcal{R}\mathcal{G}_0 \textstyle\sum_{k=1}^K \alpha_k^2 \pi_k\Big],
\end{aligned}
$$

Next, using Assumptions 3 and 5 and similar arguments to the proof of Lemma 3 in Section G.3, we can upper-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows:

$$
\begin{aligned}
\mathbb{E}[\mathrm{Reg}_K] &\leq \mathcal{O}\left[\mathcal{R}\mathcal{G}_0 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathbb{E}\left[\mathcal{R}\mathcal{G}_1 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 [f(\overline{x}_k) - f^*]^2}\right]\right] \\
&\overset{(a)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathbb{E}\left[\mathcal{R}\mathcal{G}_1 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 [f(x_k) - f^*]^2}\right]\right] \\
&\quad + \mathcal{O}\left[\mathbb{E}\left[\mathcal{R}\mathcal{G}_1 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 (\mathcal{G}_0 + \mathcal{G}_1[f(x_k) - f^*])^2 \exp(2\mathcal{G}_1 \|\overline{x}_k - x_k\|_2) \|\overline{x}_k - x_k\|_2^2}\right]\right] \\
&\overset{(b)}{\leq} \mathcal{O}\left[\mathcal{R}\mathcal{G}_0 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathbb{E}\left[\mathcal{R}\mathcal{G}_1 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 [f(x_k) - f^*]^2}\right]\right] \\
&\quad + \mathcal{O}\left[\mathbb{E}\left[\frac{\mathcal{R}}{T} \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 (\mathcal{G}_0 + \mathcal{G}_1[f(x_k) - f^*])^2}\right]\right] \\
&\leq \mathcal{O}\left[\mathcal{R}\mathcal{G}_0 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2} + \mathbb{E}\left[\mathcal{R}\mathcal{G}_1 \sqrt{\textstyle\sum_{k=1}^K \alpha_k^2 \pi_k^2 [f(x_k) - f^*]^2}\right]\right],
\end{aligned}
$$

where (a) uses Assumption 5, Lemma 9, and Lemma 1; (b) uses the upper bound on $\|\overline{x}_k - x_k\|_2$ above. Moreover, using similar arguments to the proof of Lemma 5 in Section G.7, with the modified

weight recursion in eq. (51), and observing that the blockwise bound $\pi_k \leq \mathcal{O}[\pi_{k_s}]$ still holds, we can further upper-bound $\mathbb{E}[\mathrm{Reg}_K]$ as follows:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_K] &\leq \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \sqrt{T}\mathcal{R}^2\mathcal{G}_0\mathcal{G}_1\textstyle\sum_{k=1}^K \alpha_k^2\pi_k + \frac{\mathcal{R}\mathcal{G}_1}{\sqrt{T}}\textstyle\sum_{k=1}^K \alpha_k\pi_k\mathbb{E}[f(x_k)-f^*]\right] \\ &\overset{(a)}{\leq} \tfrac{1}{4}\gamma\textstyle\sum_{k=1}^K \alpha_k\pi_k\mathbb{E}[f(x_k)-f^*] + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \sqrt{T}\mathcal{R}^2\mathcal{G}_0\mathcal{G}_1\textstyle\sum_{k=1}^K \alpha_k^2\pi_k\right]\end{aligned}$$

where (a) uses the definition of T. Now, we combine this with the previously obtained upper bound on $\pi_K\mathbb{E}[f(x_K)-f^*]$ and get the following inequality:

$$\begin{aligned}\pi_K\mathbb{E}[f(x_K)-f^*] \leq [f(x_0)-f^*] + \tfrac{1}{2}\gamma\textstyle\sum_{k=1}^K \alpha_k\pi_k\mathbb{E}[f(x_k)-f^*] \\ + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{k=1}^K \alpha_k^2\pi_k^2} + \sqrt{T}\mathcal{R}^2\mathcal{G}_0\mathcal{G}_1\textstyle\sum_{k=1}^K \alpha_k^2\pi_k\right],\end{aligned}$$

which holds for all $K \geq T$. Finally, let $\Delta_k = \mathbb{E}[f(x_k)-f^*]$, and let δ_k be defined as follows:

$$\delta_k = e\mathcal{F} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{i=1}^k \alpha_i^2\pi_i^2} + T\mathcal{R}\mathcal{G}_0\textstyle\sum_{i=1}^k \alpha_i^2\pi_i\right].$$

Using the previous inequality and similar arguments to the proof of Theorem 6 in Section G.6, for all $K \in \mathbb{N}$, we obtain the following inequality:

$$\pi_K\Delta_K \leq \delta_K + \tfrac{1}{2}\gamma\textstyle\sum_{k=1}^K \alpha_k\pi_k\Delta_k.$$

Next, for $k \leq ST$, we can upper-bound δ_k as follows:

$$\begin{aligned}\delta_k &\overset{(a)}{=} e\mathcal{F} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{i=1}^k \alpha_1^2(1-\gamma\alpha_1)^{-2i}} + T\mathcal{R}\mathcal{G}_0\textstyle\sum_{i=1}^k \alpha_1^2(1-\gamma\alpha_1)^{-i}\right] \\ &\leq e\mathcal{F} + \mathcal{O}\left[\sqrt{\alpha_1/\gamma}\pi_k\mathcal{R}\mathcal{G}_0 + T\mathcal{R}\mathcal{G}_0[\alpha_1/\gamma]\pi_k\right] \\ &\overset{(b)}{\leq} e\mathcal{F} + \mathcal{O}[\pi_k\mathcal{G}_0/[\gamma\mathcal{G}_1]],\end{aligned}$$

where (a) uses the definitions in eqs. (51) and (52); (b) uses the definition in eq. (52). Furthermore, for $k \geq ST+1$, we can upper-bound δ_k as follows:

$$\begin{aligned}\delta_k &\overset{(a)}{\leq} \delta_{ST} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\textstyle\sum_{i=1}^{k-ST} \alpha_{ST+1}^2\pi_{ST}^2(1-\gamma\alpha_{ST+1})^{-2i}}\right] \\ &\quad + \mathcal{O}\left[T\mathcal{R}\mathcal{G}_0\textstyle\sum_{i=1}^{k-ST} \alpha_{ST+1}^2\pi_{ST}(1-\gamma\alpha_{ST+1})^{-i}\right] \\ &\leq \delta_{ST} + \mathcal{O}\left[\mathcal{R}\mathcal{G}_0\sqrt{\alpha_{ST+1}/\gamma}\pi_k + T\mathcal{R}\mathcal{G}_0[\alpha_{ST+1}/\gamma]\pi_k\right] \\ &\overset{(b)}{\leq} \delta_{ST} + \mathcal{O}[\epsilon\pi_k]\end{aligned}$$

where (a) uses the definitions in eqs. (51) and (52); (b) uses the definition in eq. (52). Finally, we can upper-bound Δ_K as follows:

$$\begin{aligned}\Delta_K &\overset{(a)}{\leq} \mathcal{O}\left[\tfrac{\delta_K}{\pi_K} + \textstyle\sum_{k=1}^K \frac{\gamma\alpha_k\delta_k}{\sqrt{\pi_K\pi_{k-1}}}\right] \\ &\overset{(b)}{=} \mathcal{O}\left[\tfrac{\delta_K}{\pi_K} + \textstyle\sum_{k=1}^{ST} \frac{\gamma\alpha_k\delta_k}{\sqrt{\pi_K\pi_{k-1}}} + \sum_{k=ST+1}^K \frac{\gamma\alpha_k\delta_k}{\sqrt{\pi_K\pi_{k-1}}}\right] \\ &\overset{(c)}{\leq} \mathcal{O}\left[\tfrac{\mathcal{F}}{\pi_K} + \tfrac{\mathcal{G}_0\pi_{ST}}{\gamma\mathcal{G}_1\pi_K} + \epsilon + \textstyle\sum_{k=1}^{ST}\left[\frac{\gamma\alpha_k\mathcal{F}}{\sqrt{\pi_K\pi_{k-1}}} + \frac{\gamma\mathcal{G}_0\alpha_k\pi_k}{\gamma\mathcal{G}_1\sqrt{\pi_K\pi_{k-1}}}\right]\right] \\ &\quad + \mathcal{O}\left[\textstyle\sum_{k=ST+1}^K\left[\frac{\gamma\alpha_k\mathcal{F}}{\sqrt{\pi_K\pi_{k-1}}} + \frac{\gamma\mathcal{G}_0\alpha_k\pi_{ST}}{\gamma\mathcal{G}_1\sqrt{\pi_K\pi_{k-1}}} + \frac{\gamma\epsilon\alpha_k\pi_k}{\sqrt{\pi_K\pi_{k-1}}}\right]\right] \\ &\overset{(d)}{\leq} \mathcal{O}\left[\tfrac{\mathcal{F}}{\pi_K} + \tfrac{\mathcal{G}_0\pi_{ST}}{\gamma\mathcal{G}_1\pi_K} + \epsilon + \tfrac{\gamma\mathcal{F}\alpha_1}{\sqrt{\pi_K}}\textstyle\sum_{k=0}^{ST-1}[1-\gamma\alpha_1]^{k/2} + \frac{\gamma\mathcal{G}_0\alpha_1}{\gamma\mathcal{G}_1\sqrt{\pi_K}}\sum_{k=0}^{ST-1}[1-\gamma\alpha_1]^{-k/2-1}\right] \\ &\quad + \mathcal{O}\left[\left[\tfrac{\gamma\alpha_{ST+1}\mathcal{F}}{\sqrt{\pi_K\pi_{ST}}} + \tfrac{\gamma\mathcal{G}_0\alpha_{ST+1}\sqrt{\pi_{ST}}}{\gamma\mathcal{G}_1\sqrt{\pi_K}}\right]\textstyle\sum_{k=0}^{K-ST-1}[1-\gamma\alpha_{ST+1}]^{k/2}\right]\end{aligned}$$

$$
\begin{aligned}
&+ \mathcal{O}\Big[\tfrac{\gamma\epsilon\alpha_{ST+1}\sqrt{\pi_{ST}}}{2\sqrt{\pi_K}}\textstyle\sum_{k=0}^{K-ST-1}[1-\gamma\alpha_{ST+1}]^{-k/2-1}\Big] \\
&\overset{(e)}{\leq} \mathcal{O}\Big[\mathcal{F}[1-\gamma\alpha_1]^{ST/2}[1-\gamma\alpha_{ST+1}]^{[K-ST]/2} + [\mathcal{G}_0/[\gamma\mathcal{G}_1]][1-\gamma\alpha_{ST+1}]^{[K-ST]/2} + \epsilon\Big] \\
&\overset{(f)}{\leq} \mathcal{O}[\epsilon],
\end{aligned}
$$

where (a) uses Lemma 11 with $p = 1/2$; (b) uses the fact that $K \geq TS + 1$; (c) uses the inequalities above; (d) and (e) use the definitions in eqs. (51) and (52); (f) uses the definitions in eq. (53). □