Universal Adaptive Proximal Gradient Methods via Gradient Mapping Accumulation

Zimeng Wang & **Alp Yurtsever**
Department of Mathematics and Mathematical Statistics
Umeå University, Sweden
{zimeng.wang, alp.yurtsever}@umu.se

Abstract

We propose an adaptive proximal gradient method for minimizing the sum of two functions, where one is a simple convex function, and the other belongs to one of the three classes: nonconvex smooth, convex nonsmooth, or convex smooth. The key feature of the method is an adaptive step size that accumulates historical gradient mapping norms in the denominator. Without any modification or knowledge of problem parameters, the method converges across all three problem classes under mild bounded-iterates and bounded-variance assumptions, with rates matching those of the proximal gradient method up to logarithmic factors, in both deterministic and stochastic settings. For the convex setting, we further propose an accelerated variant. It retains a similar near-optimal convergence rate for the nonsmooth case and achieves an improved rate of order $\widetilde{O}\big(1/t^2 + \sigma/\sqrt{t}\big)$ for the smooth case, which is optimal up to logarithmic factors. Notably, we develop new techniques for controlling the effect of stochastic noise, which are applicable across all three problem classes in the stochastic setting and enable a simplified analysis.

1 Introduction

We study the following nonsmooth composite optimization problem:

$$\min_{x\in\mathbb{R}^d} F(x) := f(x) + h(x), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is either nonconvex smooth or convex (possibly nonsmooth), and $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex, proper, and lower semicontinuous, but possibly nonsmooth. Problems of this form are ubiquitous in machine learning, signal processing, and statistics. Representative examples include LASSO [39], support vector machines [7], regularized logistic regression [4], and trace-norm matrix completion [5]. In the stochastic setting, we are particularly interested in the case $f(x) = \mathbb{E}_{\xi\in\mathcal{D}}[\ell(x;\xi)]$, where $\ell(\cdot;\xi) : \mathbb{R}^d \to \mathbb{R}$ is the loss function associated with the sample ξ drawn from the underlying data space $\mathcal{D}$.

A widely used algorithm for solving (1.1) is the proximal gradient (PG) method [27, 41], which iteratively updates (see (2.1) for the definition of $\mathrm{prox}_h(\cdot)$):

$$x_{k+1} = \mathrm{prox}_{\eta_k h}(x_k - \eta_k \nabla f(x_k)) = \arg\min_{x\in\mathbb{R}^d} \Big\{ \langle \nabla f(x_k), x\rangle + \frac{1}{2\eta_k}\|x - x_k\|^2 + h(x) \Big\}, \tag{PG}$$

with a step size $\eta_k > 0$. When f is L-smooth, PG with $\eta_k = 1/L$ achieves $O(1/t)$ convergence for both nonconvex and convex f [13], with accelerated variants [3, 30, 32] improving the convex rate to the optimal $O(1/t^2)$. For convex Lipschitz f, PG achieves the optimal $O(1/\sqrt{t})$ rate [31] with decreasing step sizes $\eta_k = O(1/\sqrt{k})$. A key limitation of PG is that step sizes must be tuned to the problem class and parameters (e.g., L) to achieve the desired rates, which is often impractical.

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

A principled remedy is to use adaptive step sizes that adjust automatically without prior knowledge of the problem class or parameters. AdaGrad [8] pioneers this by accumulating $\|\nabla f(x_k)\|^2$ in the denominator of step sizes, and has inspired a rich family of adaptive methods [19, 37, 40, 46]. In the non-composite setting ($h = 0$), AdaGrad-type methods have been rigorously analyzed for both convex [9, 16, 25, 26] and nonconvex objectives [2, 17, 42, 44], automatically matching the rates of gradient methods without prior knowledge of problem parameters.

While adaptive step sizes are well-studied for non-composite problems, their extension to (1.1) remains largely unexplored. A core difficulty is that $\nabla f(x^*)$ at a stationary point x^* is generally nonzero, so accumulating $\|\nabla f(x_k)\|^2$ in an AdaGrad manner drives step sizes to zero. While this gives the optimal $O(1/\sqrt{t})$ rate for nonsmooth f [8], for smooth f it conflicts with the standard constant step size of order $1/L$. This motivates the following question:

Can we design a universal[1] *adaptive proximal gradient method for solving* (1.1) *that automatically achieves convergence rates matching those of PG for f being nonconvex smooth, convex smooth, or convex nonsmooth, without any prior knowledge of the problem-related parameters?*

1.1 Our Contributions

We answer affirmatively by proposing Algorithm 1, whose key idea is to replace squared gradient norms in the AdaGrad denominator with squared *gradient mapping* norms, which vanish at stationary points of (1.1). We show that Algorithm 1 is *universal*: it applies to three problem classes without modification or prior knowledge of parameters (e.g., smoothness constant, diameter D, or noise level σ); see also Table 1. We summarize the main contributions as follows:

- For nonconvex smooth f, Algorithm 1 achieves an $O(1/t)$ stationarity rate deterministically and $\widetilde{O}\big(\frac{1}{t} + \frac{\sigma^2}{t}\sum_{k=1}^{t}\frac{1}{b_k}\big)$ with high probability, matching the best known rates of PG [11, 13]. To the best of our knowledge, this is the first convergence analysis of an AdaGrad-type method for nonconvex composite objectives.
- The same algorithm adapts to convex settings without any modification. Specifically, the averaged iterate of Algorithm 1 with a unit batch size converges to an optimal solution at a rate of $\widetilde{O}\big((1+\sigma)/\sqrt{t}\big)$ for Lipschitz f, and improves to $\widetilde{O}(1/t + \sigma/\sqrt{t})$ for smooth f. The corresponding deterministic rates follow directly by setting $\sigma = 0$.
- For convex f, we develop an accelerated variant Algorithm 2 that achieves better convergence rates when the diameter D is known. In particular, Algorithm 2 with a unit batch size achieves an $\widetilde{O}((1+\sigma)/\sqrt{t})$ rate for Lipschitz f and $\widetilde{O}(1/t^2 + \sigma/\sqrt{t})$ for smooth f. Both rates are optimal up to logarithmic factors [20].
- Furthermore, for convex (L_ν, ν)-Hölder-smooth objectives with $\nu \in [0, 1]$, the two algorithms also interpolate between the Lipschitz-continuous and smooth rates. Specifically, their expected objective gaps are $\widetilde{O}(t^{-(1+\nu)/2} + \sigma/\sqrt{t})$ for Algorithm 1 and $\widetilde{O}(t^{-(1+3\nu)/2} + \sigma/\sqrt{t})$ for Algorithm 2. See Theorems 7 and 8 in Appendix F.

A key technical challenge in the stochastic setting is that the adaptive step sizes become random variables, making it difficult to bound terms involving their products with other random variables. We overcome this difficulty by estimating cumulative inner products between the stochastic noise and iterate differences, which enables us to establish convergence with simplified analysis across different problem classes. Additionally, our analysis removes the restrictive assumption that the stochastic gradients are uniformly bounded almost surely, which is frequently imposed in the related literature for nonsmooth analysis (see, e.g., [16, 17, 25, 26, 44]).

The remainder of the paper is organized as follows. Section 2 introduces the problem setup and Algorithm 1. Sections 3 and 4 present convergence results for nonconvex and convex settings, respectively. Section 5 introduces the accelerated variant Algorithm 2 and establishes its convergence in the convex setting. Section 6 presents numerical experiments, and Section 7 concludes the paper. All technical proofs are deferred to Appendices A–D. Appendix E discusses the boundedness assumption and a block-diagonal extension. Appendix F gives the stochastic Hölder-smooth extensions, and Appendix G provides additional experiments.

[1] Here, *universal* means that one method covers the three problem classes. The notion of universality in Nesterov [33] means adaptation to an unknown Hölder smoothness level, which we address in Appendix F for convex f.

Table 1: Summary of convergence results for different classes of f. "Full" means full batch (deterministic). $\widetilde{O}(\cdot)$ hides constants and log factors.

Algorithm	Problem Class f			Batch	Convergence	Assumptions	Ref.
	Convex	Smooth	Lipschitz				
Alg. 1	×	✓	×	Full	$O(1/t)$	None	Thm. 1
	×	✓	×	b_k	$\widetilde{O}\big(\frac{1}{t}+\frac{\sigma^2}{t}\sum_k \frac{1}{b_k}\big)$	Asm. 1(a),2	Thm. 2
	✓	×	✓	Full	$\widetilde{O}(1/\sqrt{t})$	Asm. 1(b)	Cor. 1
	✓	×	✓	1	$\widetilde{O}((1+\sigma)/\sqrt{t})$	Asm. 1(b),2	Thm. 3
	✓	✓	×	Full	$O(1/t)$	Asm. 1(b)	Cor. 2
	✓	✓	×	1	$\widetilde{O}(1/t+\sigma/\sqrt{t})$	Asm. 1(b),2	Thm. 4
Alg. 2	✓	×	✓	Full	$\widetilde{O}(1/\sqrt{t})$	Asm. 3; $\eta>\frac{\sqrt{2}}{2}D$	Cor. 3
	✓	×	✓	1	$\widetilde{O}((1+\sigma)/\sqrt{t})$	Asm. 2,3; $\eta>\frac{\sqrt{2}}{2}D$	Thm. 5
	✓	✓	×	Full	$O(1/t^2)$	Asm. 3; $\eta>\frac{\sqrt{2}}{2}D$	Cor. 4
	✓	✓	×	1	$\widetilde{O}(1/t^2+\sigma/\sqrt{t})$	Asm. 2,3; $\eta>\frac{\sqrt{2}}{2}D$	Thm. 6

1.2 Related Work

Proximal Gradient Methods. The PG method (PG) traces back to the splitting algorithms of Lions and Mercier [27] and Passty [36], and is surveyed comprehensively in Parikh and Boyd [35]. Special cases include projected gradient descent [24] (indicator h) and ISTA/FISTA [3] (ℓ_1 regularizer h). For convex smooth f, the $O(1/t)$ rate is classical and can be accelerated to the optimal $O(1/t^2)$ via Nesterov-type momentum [3, 30, 32]. For convex nonsmooth f, PG achieves the optimal $O(1/\sqrt{t})$ rate [31]. Nesterov's universal gradient method [33] achieves near-optimal rates for both smooth and nonsmooth objectives through line search. In the stochastic setting, Lan [20] establishes optimal rates for convex objectives. These results are later extended to nonconvex cases [12, 13].

Adaptive Gradient Methods. AdaGrad [8] introduced adaptive step sizes by accumulating squared gradient norms, with popular variants including Adam [19] and AMSGrad [37]. In the non-composite setting ($h = 0$) or the constrained setting (h is the indicator function of a convex set), universal accelerated methods [9, 16, 25, 26] improve over AdaGrad by achieving near-optimal rates simultaneously for convex smooth and nonsmooth objectives. For nonconvex objectives, Kavis et al. [17] and Ward et al. [44] provide high-probability guarantees under a bounded-gradient assumption, which is later removed through refined analysis [2, 10, 42]. Extensions to the composite setting (1.1) are less studied and mainly focus on the convex regime. The original AdaGrad [8] handles the composite structure but may converge slowly for smooth f due to the accumulation of squared gradient norms. This issue is later addressed by Joulani et al. [15] and Rodomanov et al. [38], who accumulate norms of gradient differences instead. Nevertheless, these works are restricted to the convex regime, and their step sizes use an estimate of the diameter D, as do accelerated adaptive methods [9, 16, 26] and our Algorithm 2, but not our Algorithm 1. More closely related to our work are AdaGrad+ and AdaACSA [9] for constrained convex problems, whose step sizes are driven by gradient-mapping-type quantities: when h is the indicator function of a convex set, Algorithms 1 and 2 reduce to their scalar versions up to the choice of η (see also [9, Remark 5.5]). Compared with these methods, our work provides several improvements: (i) general composite problems; (ii) one method for all three problem classes, including the nonconvex one; (iii) guarantees for Algorithm 1 with any $\eta > 0$, whereas both methods in [9] set η proportional to D; and (iv) a simpler analysis. Gradient mappings are also accumulated in the adaptive bilevel method of Antonakopoulos et al. [1], which does not target universality across problem classes. For nonconvex composite problems, we are only aware of Yun et al. [45], who propose a general adaptive PG framework. However, they only provide a rough convergence analysis for non-increasing step sizes, which requires prior knowledge of the smoothness constant, making the method not fully adaptive. A separate line of adaptivity, pioneered by Malitsky and Mishchenko [28], uses step sizes based on local gradient variation, achieving adaptivity to local smoothness without line search. This approach is later extended to the composite setting [21, 22, 29].

2 Problem Setup

Notation and Definitions. We use $\|\cdot\|$ for the Euclidean norm, $[a]_+ := \max\{0, a\}$ for the positive part, and $\log(\cdot)$ for the natural logarithm. For a stochastic algorithm, we denote by $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$ the conditional expectation given the σ-algebra $\mathcal{F}_{k-1}$, which is the expectation given all the information available at step k. We use ∇f to denote the gradient or a subgradient of f.

For a proper, lower-semicontinuous, and convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we denote its (effective) domain by $\mathrm{dom}(h) := \{x \in \mathbb{R}^d : h(x) < +\infty\}$ and its proximal operator by $\mathrm{prox}_h(\cdot)$, defined as

$$\mathrm{prox}_h(x) := \operatorname*{arg\,min}_{z \in \mathbb{R}^d} \Big\{ h(z) + \frac{1}{2}\|z - x\|^2 \Big\}, \ \forall x \in \mathbb{R}^d. \tag{2.1}$$

The proximal operator $\mathrm{prox}_h(\cdot)$ is nonexpansive:

$$\|\mathrm{prox}_h(x) - \mathrm{prox}_h(y)\| \le \|x - y\|, \ \forall x, y \in \mathbb{R}^d. \tag{2.2}$$

Throughout the paper, we always assume that the objective function F in (1.1) is bounded from below, and we denote $F^* := \inf_{x \in \mathbb{R}^d} F(x)$. If F is convex, we also assume that it attains a global minimizer, which we denote by $x^* \in \arg\min_{x \in \mathbb{R}^d} F(x)$, so that $F^* = F(x^*)$. When f is nonconvex but differentiable, stationarity is measured via the gradient mapping [13]

$$\mathcal{G}_\eta(x) := \frac{1}{\eta}\big(x - \mathrm{prox}_{\eta h}(x - \eta \nabla f(x))\big), \tag{2.3}$$

which satisfies $\mathcal{G}_\eta(x^*) = 0$ if and only if x^* is a stationary point of (1.1).

Algorithm Design. Recall that AdaGrad [8] employs adaptive step sizes that accumulate $\|\nabla f(x_k)\|^2$ in the denominator. The motivation is that one has $\nabla f(x^*) = 0$ for any stationary point x^* when $h = 0$. Consequently, for smooth f, the accumulated norms $\sum_{i=1}^k \|\nabla f(x_i)\|^2$ grow slowly as $\nabla f(x_k)$ diminishes, making the step size approximately constant. For nonsmooth f, the subgradient norms may stay bounded away from zero, leading to decreasing step sizes $O(1/\sqrt{k})$.

However, this strategy does not directly extend to the composite setting. The reason is that the gradient $\nabla f(x^*)$ at a stationary point x^* of (1.1) is generally nonzero since the first-order optimality condition merely requires $-\nabla f(x^*) \in \partial h(x^*)$. Hence, accumulating $\|\nabla f(x_k)\|^2$ would drive the step size to zero even for smooth f, which is undesirable. The appropriate stationarity measure in the composite setting is the gradient mapping $\mathcal{G}_\eta(\cdot)$ (2.3), which reduces to ∇f if $h = 0$ and satisfies $\mathcal{G}_\eta(x^*) = 0$ at any stationary point x^* of (1.1). Moreover, the update rules in the form of (PG) provide direct access to the gradient mapping $\mathcal{G}_k := \mathcal{G}_{\eta_k}(x_k)$ due to

$$\mathcal{G}_k = \frac{1}{\eta_k}\big(x_k - \mathrm{prox}_{\eta_k h}(x_k - \eta_k \nabla f(x_k))\big) = \frac{1}{\eta_k}(x_k - x_{k+1}). \tag{2.4}$$

This motivates us to accumulate $\|\mathcal{G}_k\|^2$ in the denominator of step sizes, yielding the following adaptive scheme at iteration k:

$$\eta_k = \frac{\eta}{S_k}, \quad S_k^2 = \gamma^2 + \sum_{j=1}^{k-1} \|\mathcal{G}_j\|^2, \tag{2.5}$$

where $\gamma, \eta > 0$ are hyperparameters. Unlike standard AdaGrad, we accumulate up to index $k-1$ instead of k because $\mathcal{G}_k = (x_k - x_{k+1})/\eta_k$ is not available until the k-th iteration is performed. Note that the step size (2.5) is different from the ones in Joulani et al. [15] and Rodomanov et al. [38], which accumulate squared norms of gradient differences to handle the composite structure. We focus on scalar step sizes for notational simplicity; see Appendix E.2 for a block-diagonal extension.

Applying (2.5), we obtain the adaptive proximal gradient method for solving (1.1), as summarized in Algorithm 1. For stochastic settings, we define the stochastic gradient mapping as

$$\widetilde{\mathcal{G}}_k := \frac{1}{\eta_k}\big(x_k - \mathrm{prox}_{\eta_k h}(x_k - \eta_k g_k)\big) = \frac{1}{\eta_k}(x_k - x_{k+1}). \tag{2.6}$$

Then we similarly have $S_k^2 = \gamma^2 + \sum_{j=1}^{k-1} \|\widetilde{\mathcal{G}}_j\|^2$.

Algorithm 1 Adaptive Proximal Gradient Method for Solving (1.1)

Require: $x_1 \in \mathrm{dom}(h)$; hyperparameters $\gamma, \eta > 0$; batch sizes $\{b_k\}_{k\geq 1}$ (stochastic setting only).
  Initialize $S_1 := \gamma$.
  for $k = 1, 2, \ldots$ **do**
    Compute $g_k := \nabla f(x_k)$ (deterministic) or $g_k := \frac{1}{b_k}\sum_{j=1}^{b_k} \nabla \ell(x_k; \xi_{k_j})$ (stochastic) for randomly sampled $\{\xi_{k_j}\}_{j=1}^{b_k}$.
    Update $x_{k+1} := \arg\min_{x\in\mathbb{R}^d} \Big\{ \langle g_k, x\rangle + \frac{S_k}{2\eta}\|x - x_k\|^2 + h(x) \Big\}$.
    Update $S_{k+1}^2 = S_k^2\big(1 + \|x_{k+1} - x_k\|^2/\eta^2\big)$.
  end for

To analyze Algorithm 1, we make the following boundedness assumption on the iterates $\{x_k\}_{k\geq 1}$. Similar boundedness assumptions are widely adopted for the analysis of AdaGrad-type methods in the literature (see, e.g., [8, 9, 15, 16]). In the context of the composite problem (1.1), this assumption holds if, e.g., $\mathrm{dom}(h)$ is bounded or the operator $\arg\min_{x\in\mathbb{R}^d}\{\langle u, x\rangle + h(x)\}$ is bounded for all $u \in \mathbb{R}^d$ [12, Lemma 2].

Assumption 1. There exists a constant $D > 0$ such that (a) $\|x_{k+1} - x_k\| \leq D$ for all $k \geq 1$ (bounded steps), and, if F is convex, (b) $\|x_k - x^*\| \leq D$ for all $k \geq 1$ (bounded iterates).

Note that while Assumption 1 is used in the analysis, knowledge of the diameter D is not required for implementing Algorithm 1. Part (a), which holds automatically for smooth f in the deterministic setting (see Appendix E), is needed due to the delayed accumulation in (2.5). Since part (b) implies part (a) with $2D$ in place of D, our results require only part (a) for nonconvex f and only part (b) for convex f. For notational convenience, we define $\Delta_k := F(x_k) - F^*$.

3 Convergence Analysis in the Nonconvex Setting

In this section, we study the convergence of Algorithm 1 when f is L-smooth but possibly nonconvex. For ease of presentation, we use $\lesssim$ to hide constants depending only on fixed quantities (e.g., L, G, γ, η, D) hereafter, i.e., $A \lesssim B$ if there exists a constant $c > 0$ such that $A \leq cB$.

Deterministic Setting. We first consider $g_k := \nabla f(x_k)$. The following theorem, whose proof relies on the boundedness of $\{\Delta_k\}_{k\geq 1}$, establishes an $O(1/t)$ stationarity rate for Algorithm 1.

Theorem 1. *Suppose f is L-smooth. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $g_k := \nabla f(x_k)$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have $\frac{1}{t}\sum_{k=1}^{t} \|\mathcal{G}_k\|^2 \lesssim \frac{1}{t}$.*

Stochastic Setting. We proceed to study the case $g_k := \frac{1}{b_k}\sum_{j=1}^{b_k} \nabla \ell(x_k; \xi_{k_j})$ for some batch size $b_k \geq 1$, where the samples $\{\xi_{k_j}\}$ are drawn independently of each other and of the past iterates. We first make the following unbiasedness and bounded-variance assumption.

Assumption 2. $\mathbb{E}_k[\nabla \ell(x_k; \xi)] = \nabla f(x_k)$ for all $k \geq 1$, where the expectation is taken w.r.t. the randomness of ξ. Moreover, there exists $\sigma \geq 0$ such that $\mathbb{E}_k[\|\nabla \ell(x_k; \xi) - \nabla f(x_k)\|^2] \leq \sigma^2$, $\forall k \geq 1$.

Analysis of Algorithm 1 in such a stochastic setting is more challenging, because the adaptive step sizes $\{\eta_k\}_{k\geq 1}$ become random variables. A direct consequence is that our deterministic analysis relying on bounding objective gaps is not directly applicable. To see this, while we may still show that the objective gaps $\{\Delta_k\}_{k\geq 1}$ are uniformly bounded in expectation, it is hard to bound

$$\mathbb{E}\Big[\sum_{k=2}^{t}\Big(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\Big)\Delta_k\Big]$$

due to the coupling between the two random variables $\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}$ and, which is crucial for establishing Theorem 1. This issue regarding randomness in $\{\eta_k\}_{k\geq 1}$ is also identified as the main bottleneck for nonconvex analysis of AdaGrad in non-composite settings [2, 10, 42, 44]. Most of these works handle it by introducing step-size surrogates that are measurable given all historical iterates, which leads to highly involved analyses. Besides, the composite structure of (1.1) introduces additional

challenges because in this case $\langle \nabla f(x_k), x_{k+1} - x_k \rangle = -\eta_k \langle \nabla f(x_k), \widetilde{\mathcal{G}}_k \rangle \neq -\eta_k \langle \nabla f(x_k), g_k \rangle$, which makes existing nonconvex analyses for non-composite settings inapplicable.

To overcome this difficulty, we develop a new approach by carefully estimating the sum of inner products between the stochastic noise and iterate differences, which allows us to bound $\mathbb{E}[S_t]$ without introducing new surrogates and leads to a substantially simplified analysis. We present this technical ingredient in the following lemma. Note that it does not rely on any conditions on f and hence is applicable across different settings of f.

Lemma 1. *Suppose Assumptions 1(a) and 2 hold. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with batch sizes $\{b_k\}_{k\geq 1}$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have*

$$\mathbb{E}\Big[\sum_{k=1}^{t} \langle \nabla f(x_k) - g_k, x_{k+1} - x_k \rangle\Big] \leq \sigma \sqrt{2(\eta^2 + D^2) \sum_{k=1}^{t} \frac{1}{b_k}} \sqrt{\log \mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]}.$$

Using Lemma 1, we can establish convergence to stationary points for Algorithm 1.

Theorem 2. *Suppose f is L-smooth and Assumptions 1(a) and 2 hold. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with batch sizes $\{b_k\}_{k\geq 1}$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have*

$$\left(\mathbb{E}\left[\sqrt{\frac{1}{t}\sum_{k=1}^{t} \|\mathcal{G}_k\|^2}\right]\right)^2 \lesssim \frac{1}{t}\left(1 + \sigma^2 \sum_{k=1}^{t} \frac{1}{b_k}\right)\left(1 + \log^2\left(1 + \sigma\sqrt{\sum_{k=1}^{t} \frac{1}{b_k}}\right)\right). \tag{3.1}$$

Remark 1. With unit batch size $b_k \equiv 1$, the bound in Theorem 2 reduces to $\widetilde{O}\big(\frac{1}{t} + \sigma^2\big)$, which only ensures convergence to a neighborhood of a stationary point. This marks a key difference from AdaGrad analysis in the non-composite setting, where unit batch size suffices for convergence to a stationary point [2, 10, 42, 44]. Such behavior is not an artifact of our analysis, as it matches the behavior of PG in the stochastic nonconvex setting [13], whose convergence bound also contains a residual term proportional to $\frac{\sigma^2}{t}\sum_{k=1}^{t}\frac{1}{b_k}$. To ensure convergence to a stationary point, one may take increasing batch sizes, e.g., $b_k = \Omega(\sigma\sqrt{k})$, which yields the rate $\widetilde{O}\big(\frac{1}{t} + \frac{\sigma}{\sqrt{t}}\big)$, matching the result of Ghadimi et al. [13].

Remark 2. Applying Markov's inequality to Theorem 2 further gives a high-probability bound: for any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$, $\frac{1}{t}\sum_{k=1}^{t} \|\mathcal{G}_k\|^2$ is at most $1/\epsilon^2$ times the right-hand side of (3.1) (see Appendix B.3).

4 Convergence Analysis in the Convex Setting

In this section, we study the convergence of Algorithm 1 under the condition that f is convex. Leveraging Lemma 1, we show that Algorithm 1 automatically achieves convergence rates matching those of PG for both Lipschitz continuous and smooth f. For brevity, we focus on stochastic settings and fix $b_k \equiv 1$ (i.e., $g_k = \nabla \ell(x_k; \xi_k)$) throughout this section, and the established results easily extend to mini-batch and deterministic settings. Denote $\bar{x}_t := \frac{1}{t}\sum_{k=1}^{t} x_{k+1}$.

Lipschitz Continuous f. We establish the following convergence result for the averaged iterate of Algorithm 1 when f is convex and G-Lipschitz continuous.

Theorem 3. *Suppose f is convex and G-Lipschitz continuous, and Assumptions 1(b) and 2 hold. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $b_k \equiv 1$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have*

$$\mathbb{E}[F(\bar{x}_t) - F^*] \lesssim \frac{(1+\sigma)\sqrt{\log\big(1 + (1+\sigma^2)t\big)}}{\sqrt{t}}.$$

Setting $\sigma = 0$ in Theorem 3, we obtain an $O\big(\sqrt{\log(1+t)/t}\big)$ deterministic rate for Algorithm 1.

Corollary 1. *Suppose f is convex and G-Lipschitz continuous, and Assumption 1(b) holds. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $g_k = \nabla f(x_k)$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, it holds that $F(\bar{x}_t) - F^* \lesssim \sqrt{\log(1+t)/t}$.*

Algorithm 2 Adaptive Accelerated Proximal Gradient Method for Solving (1.1)

Require: initial points $y_1 = z_1 \in \mathrm{dom}(h)$; hyperparameters $\gamma, \eta > 0$; batch sizes $\{b_k\}_{k\geq 1}$ (stochastic setting only).
Initialize $S_1 := \gamma$ and $\alpha_0 := 0$.
for $k = 1, 2, \ldots$ **do**
Update weight $\alpha_k := \frac{1+\sqrt{1+4\alpha_{k-1}^2}}{2}$; set $\theta_k := 1/\alpha_k$.
Set $x_k = (1-\theta_k)y_k + \theta_k z_k$.
Compute $g_k := \nabla f(x_k)$ (deterministic) or $g_k := \frac{1}{b_k}\sum_{j=1}^{b_k} \nabla \ell(x_k; \xi_{k_j})$ (stochastic) for randomly sampled $\{\xi_{k_j}\}_{j=1}^{b_k}$.
Update the next iterates y_{k+1} and z_{k+1} via

$$z_{k+1} = \arg\min_{x\in\mathbb{R}^d} \Big\{ \langle g_k, x\rangle + \frac{\theta_k S_k}{2\eta}\|x - z_k\|^2 + h(x) \Big\}, \tag{5.1}$$

$$y_{k+1} = x_k + \theta_k\big(z_{k+1} - z_k\big). \tag{5.2}$$

Update $S_{k+1}^2 = S_k^2(1 + \|z_{k+1} - z_k\|^2/\eta^2)$.
end for

Smooth f. When f is convex and L-smooth, we derive the following convergence result.

Theorem 4. *Suppose f is convex and L-smooth, and Assumptions 1(b) and 2 hold. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $b_k \equiv 1$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have*

$$\mathbb{E}[F(\bar{x}_t) - F^*] \lesssim \frac{1 + \log\big(1 + \sigma\sqrt{t}\big)}{t} + \frac{\sigma \log\big(1 + \sigma\sqrt{t}\big)}{\sqrt{t}}.$$

Setting $\sigma = 0$ in Theorem 4 leads to an $O(1/t)$ deterministic convergence rate for Algorithm 1.

Corollary 2. *Suppose f is convex and L-smooth, and Assumption 1(b) holds. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $g_k = \nabla f(x_k)$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have $F(\bar{x}_t) - F^* \lesssim 1/t$.*

5 Acceleration under Convexity and Known Diameter

In this section, we study an accelerated version of Algorithm 1 by combining the accelerated stochastic approximation (AC-SA) algorithm [20] with adaptive step sizes. The resulting method is summarized in Algorithm 2, which extends the scalar AdaACSA [9] to composite problems.

Since the update of the intermediate point z_{k+1} in Algorithm 2 requires solving a proximal subproblem (5.1), we replace Assumption 1 with the boundedness of $\{z_k\}_{k\geq 1}$ throughout this section.

Assumption 3. There exists $D > 0$ such that $\|z_k - x^*\| \leq D$ for all $k \geq 1$.

Under Assumption 3, we also have bounded steps since $\|z_{k+1} - z_k\| \leq \|z_{k+1} - x^*\| + \|z_k - x^*\| \leq 2D$ for all $k \geq 1$. As in the previous section, we focus on convergence results for Algorithm 2 in stochastic settings and fix the batch size $b_k \equiv 1$ for simplicity. We demonstrate that, when f is convex and the diameter D in Assumption 3 is known, Algorithm 2 preserves a near-optimal convergence rate similar to that of Algorithm 1 if f is Lipschitz continuous. Moreover, it achieves an improved convergence rate for L-smooth f, which is optimal up to logarithmic factors [20]. A further difference is that while the convergence of Algorithm 1 is measured at the averaged iterate $\bar{x}_t$, that of Algorithm 2 is at the last iterate y_t, which is preferred in practice as it requires no additional computation.

Lipschitz Continuous f. We first consider G-Lipschitz continuous f. The following theorem shows that if $\eta > \sqrt{2}D/2$, then Algorithm 2 achieves a near-optimal convergence rate $\widetilde{O}\big((1+\sigma)/\sqrt{t}\big)$.

Theorem 5. *Suppose f is convex and G-Lipschitz continuous, and Assumptions 2 and 3 hold. Let $\{y_k\}_{k\geq 1}$ be generated by Algorithm 2 with $b_k \equiv 1$, $\gamma > 0$, and $\eta > \sqrt{2}D/2$. Then for any $t \geq 1$, we have*

$$\mathbb{E}[F(y_{t+1}) - F^*] \lesssim \frac{1+\sigma}{\sqrt{t}}\sqrt{\log\big(1 + (1+\sigma^2)t^3\big)}.$$

Setting $\sigma = 0$ in Theorem 5, we obtain a deterministic rate $O\big(\sqrt{\log(1+t)/t}\big)$ for Algorithm 2.

Corollary 3. *Suppose f is convex and G-Lipschitz continuous, and Assumption 3 holds. Let $\{y_k\}_{k\geq 1}$ be generated by Algorithm 2 with $g_k = \nabla f(x_k)$, $\gamma > 0$, and $\eta > \sqrt{2}D/2$. Then for any $t \geq 1$, it holds that $F(y_{t+1}) - F^* \lesssim \sqrt{\log(1+t)/t}$.*

Smooth f. When f is convex and L-smooth, we show in the upcoming theorem that Algorithm 2 achieves a near-optimal convergence rate if $\eta > \sqrt{2}D/2$. This rate can be significantly faster than that of Algorithm 1, particularly when the noise level σ^2 is small.

Theorem 6. *Suppose f is convex and L-smooth, and Assumptions 2 and 3 hold. Let $\{y_k\}_{k\geq 1}$ be generated by Algorithm 2 with $b_k \equiv 1$, $\gamma > 0$, and $\eta > \sqrt{2}D/2$. Then for any $t \geq 1$, we have*

$$\mathbb{E}[F(y_{t+1}) - F^*] \lesssim \frac{1 + \log\left(1 + \sigma t^{3/2}\right)}{t^2} + \frac{\sigma \log\left(1 + \sigma t^{3/2}\right)}{\sqrt{t}}.$$

Setting $\sigma = 0$ in Theorem 6, we derive an optimal deterministic rate $O\big(1/t^2\big)$ for Algorithm 2.

Corollary 4. *Suppose f is convex and L-smooth, and Assumption 3 holds. Let $\{y_k\}_{k\geq 1}$ be generated by Algorithm 2 with $g_k = \nabla f(x_k)$, $\gamma > 0$, and $\eta > \sqrt{2}D/2$. Then for any $t \geq 1$, it holds that $F(y_{t+1}) - F^* \lesssim 1/t^2$.*

Remark 3. The condition $\eta > \sqrt{2}D/2$ ensures that, in our bound on $\mathbb{E}[S_{t+1}]$, the term $-\eta S_{t+1}$ from the adaptive step sizes dominates the term $\frac{D^2}{2\eta} S_{t+1}$ from the telescoping distances (see Appendix D.1). Note that accelerated AdaGrad-type methods including [9, 16, 26] also need η proportional to D. This condition $\eta > \sqrt{2}D/2$ is only for theoretical analysis. Empirically, Algorithm 2 also converges for η less than $\sqrt{2}D/2$, as shown in Appendix G.5.

Remark 4. While Algorithm 2 achieves accelerated convergence guarantees for convex objectives, whether it converges to stationary points of smooth nonconvex objectives at the optimal rate remains an open question. Establishing such guarantees while preserving the accelerated rates in the convex case is an interesting direction for future research.

6 Experiments

In this section, we conduct numerical experiments on the nonconvex support vector machine (SVM) problem [43] with ℓ_1 regularization and box constraints:

$$\min_{x\in[-R,R]^d} F(x) := \frac{1}{n}\sum_{i=1}^{n} \left[1 - \tanh\left(b_i \langle x, a_i\rangle\right)\right] + \frac{\mu}{2}\|x\|_2^2 + \lambda\|x\|_1, \tag{6.1}$$

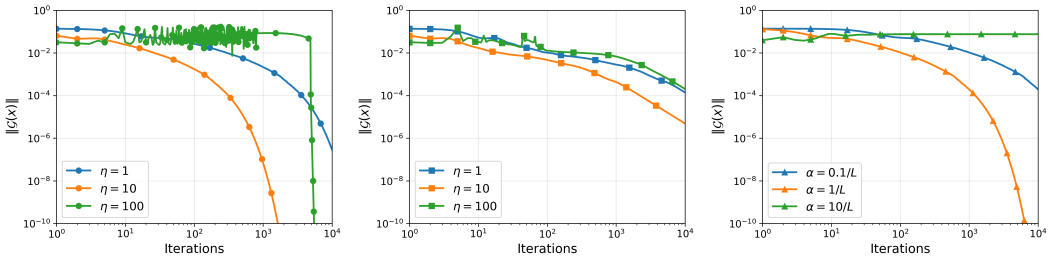
where $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$ denotes the i-th sample with feature vector $a_i \in \mathbb{R}^d$ and label $b_i \in \{-1, 1\}$. This problem fits into the composite formulation (1.1) by defining

$$f := \frac{1}{n}\sum_{i=1}^{n} \left[1 - \tanh\left(b_i \langle \cdot, a_i\rangle\right)\right] + \frac{\mu}{2}\|\cdot\|_2^2, \quad h := \lambda\|\cdot\|_1 + \iota_{[-R,R]^d}(\cdot),$$

where $\iota_{[-R,R]^d}(\cdot)$ denotes the indicator function of the set $[-R, R]^d$. Moreover, since $\mathrm{dom}(h)$ is bounded, Assumptions 1 and 3 are satisfied for Algorithms 1 and 2, respectively. All experiments were conducted on an Apple M4 Pro CPU.

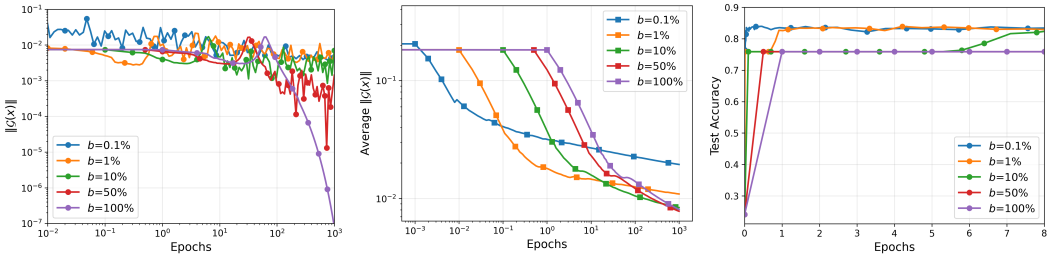
For all experiments on (6.1), we set the box parameter to $R = 50$ and the regularization parameters to $\mu = \lambda = 10^{-3}$. Since the objective function F in (6.1) is nonconvex, we measure the performance of algorithms using the norm of the gradient mapping $\mathcal{G}(\cdot) := \mathcal{G}_1(\cdot) = x - \mathrm{prox}_h\big(x - \nabla f(x)\big)$. Furthermore, due to the separability of the ℓ_1-norm and the box constraints, the proximal operator of h can be computed efficiently by first applying the soft-thresholding operator and then projecting onto the box constraints.

Comparison in Deterministic Settings. We first compare the full-batch implementations of three methods on the MNIST dataset [23]: Algorithm 1, Algorithm 2, and the proximal gradient (PG) method. We convert the dataset into a binary classification task by assigning digits 0–5 to label -1



(a) Algorithm 1 (b) Algorithm 2 (c) PG

Figure 1: Evolution of $\|\mathcal{G}(x)\|$ for solving the nonconvex SVM problem (6.1) on the MNIST dataset in the deterministic setting.



(a) $\|\mathcal{G}(x_t)\|$ (b) $\frac{1}{t}\sum_{k=1}^{t}\|\mathcal{G}(x_k)\|$ (c) Test accuracy

Figure 2: Effect of batch size on Algorithm 1 for solving the nonconvex SVM problem (6.1) on the a9a dataset. Here, $b = 1\%$ means the gradient batch size is 1% of the training set size.

and digits 6–9 to label $+1$. Our goal is to assess the robustness of these methods under different step-size schemes. For Algorithms 1 and 2, we fix $\gamma = 1$ and vary $\eta \in \{1, 10, 100\}$. For PG, we use constant step sizes $\eta_k \equiv \alpha \in \{0.1/L, 1/L, 10/L\}$ (classical theory suggests that $\alpha = 1/L$ is a safe choice [13]). Each method is run for 10,000 iterations, using the same initial point for a fair comparison.

Figure 1 shows the evolution of the gradient mapping norm for all three methods under different step-size schemes. We observe that PG converges when the step size is $1/L$ or smaller, but may diverge for larger values such as $10/L$. In contrast, Algorithm 1 exhibits fast and robust convergence to stationary points across all tested values of η, supporting Theorem 1 and highlighting the advantage of adaptive step sizes. Moreover, Algorithm 2 shows slower decay of the gradient mapping norm across all tested values of η compared to Algorithm 1, which indicates that Algorithm 2 may be less effective in the nonconvex setting (see also Remark 4).

Effect of Batch Size. We next investigate the effect of batch size on the performance of Algorithm 1 in the stochastic setting. Specifically, we use the a9a dataset [6] and randomly split it into a training set (80% of the samples) and a test set (the remaining 20%). We fix the hyperparameters $\gamma = 1$ and $\eta = 10$, and consider five constant batch sizes $b \in \{0.1\%, 1\%, 10\%, 50\%, 100\%\}$ of the training set size. For fairness, we run Algorithm 1 with each batch size for 1000 epochs (i.e., full passes over the training data), so that all configurations use the same total number of gradient evaluations.

Figure 2 illustrates the evolution of the last-iterate and averaged gradient mapping norms, as well as the test accuracy, for different batch sizes. As shown in plots (a) and (b), both the last-iterate and averaged gradient mapping norms decrease faster (in terms of epochs) when larger batch sizes are used. In contrast, plot (c) indicates that smaller batch sizes yield faster improvements in test accuracy during the early stages of training. This suggests a trade-off between optimization efficiency and generalization performance when selecting the batch size.

Additional Experiments. Appendix G reports experiments on logistic regression; comparisons with AdaGrad-Norm, coordinate-wise AdaGrad, the methods in [38], DoG [14], and DoWG [18] in both deterministic and mini-batch settings; and sensitivity tests on the hyperparameters.

7 Conclusion

We proposed an adaptive proximal gradient method for solving composite optimization problems, where one component is a simple convex function and the other belongs to one of the three classes: nonconvex smooth, convex nonsmooth, or convex smooth. The method is characterized by an adaptive step size that accumulates historical gradient mapping norms in the denominator, replacing the gradient norms used in standard AdaGrad to better accommodate the composite structure. We showed that the method converges across all three problem classes without any modification or knowledge of problem-related parameters, achieving rates matching those of proximal gradient methods up to logarithmic factors in both deterministic and stochastic settings. For the convex case, we further developed an accelerated variant that retains near-optimal rates for convex nonsmooth objectives and achieves an improved rate of $\widetilde{O}(1/t^2 + \sigma/\sqrt{t})$ for convex smooth objectives. To handle the randomness of the adaptive step sizes in the stochastic setting, we developed new technical tools to bound cumulative inner products between the stochastic noise and iterate differences, which are applicable across all three problem classes and enable simplified convergence analyses. Our experiments validate the theoretical findings and demonstrate the practical competitiveness of the proposed algorithms compared with classical proximal gradient methods.

Several directions remain open. The logarithmic factors in our convergence rates arise from bounding the growth of the denominator in the adaptive step sizes, and it is unclear whether they can be removed by a sharper analysis or a modified algorithm. Another challenging open direction is to develop an adaptive accelerated method that simultaneously achieves optimal convergence rates across all three problem classes without modifications or knowledge of problem-related parameters, thereby unifying the universality of the first algorithm we proposed with the accelerated rates of the second. Finally, further potential improvements include relaxing the boundedness assumptions beyond deterministic smooth settings, and removing the condition $\eta > \sqrt{2}D/2$ in the analysis of the accelerated method.

Acknowledgments and Disclosure of Funding

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] K. Antonakopoulos, S. Sabach, L. Viano, M. Hong, and V. Cevher. Adaptive bilevel optimization. *ACM/IMS Journal of Data Science*, 2(2):1–29, 2025. doi: 10.1145/3728478.

[2] A. Attia and T. Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pages 1147–1171. PMLR, 2023.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] J. Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365, 1944.

[5] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[8] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[9] A. Ene, H. L. Nguyen, and A. Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities. *arXiv preprint arXiv:2007.08840*, 2020.

[10] M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai, and R. Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pages 313–355. PMLR, 2022.

[11] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[12] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[13] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[14] M. Ivgi, O. Hinder, and Y. Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pages 14465–14499. PMLR, 2023.

[15] P. Joulani, A. Raj, A. György, and C. Szepesvári. A simpler approach to accelerated optimization: iterative averaging meets optimism. In *International Conference on Machine Learning*, pages 4984–4993. PMLR, 2020.

[16] A. Kavis, K. Levy, F. Bach, and V. Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[17] A. Kavis, K. Levy, and V. Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *International Conference on Learning Representations*, 2022.

[18] A. Khaled, K. Mishchenko, and C. Jin. DoWG unleashed: An efficient universal parameter-free gradient descent method. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[20] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[21] P. Latafat, A. Themelis, and P. Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. In *6th Annual Learning for Dynamics & Control Conference*, pages 197–208. PMLR, 2024.

[22] P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *Mathematical Programming*, 213 (1-2):433–471, 2025.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[24] E. Levitin and B. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[25] K. Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[26] K. Levy, A. Yurtsever, and V. Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[27] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[28] Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR, 2020.

[29] Y. Malitsky and K. Mishchenko. Adaptive proximal gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[30] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[31] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2003.

[32] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[33] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[34] F. Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

[35] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014.

[36] G. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

[37] S. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

[38] A. Rodomanov, X. Jiang, and S. Stich. Universality of AdaGrad stepsizes for stochastic optimization: Inexact oracle, acceleration and variance reduction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[39] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[40] T. Tieleman and G. Hinton. Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 2012.

[41] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[42] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Conference on Learning Theory*, pages 161–190. PMLR, 2023.

[43] X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.

[44] R. Ward, X. Wu, and L. Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21:1–30, 2020.

[45] J. Yun, A. Lozano, and E. Yang. Adaptive proximal gradient methods for structured neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[46] M. Zeiler. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

A Auxiliary Lemmas

Here, we introduce some auxiliary lemmas used throughout our analyses. These results are technical tools for solving certain classes of scalar inequalities or bounding sums of specific forms.

Lemma 2. *Let $a, b \geq 0$. If $x^2 \leq ax + b$, then $x^2 \leq a^2 + 2b$.*

Lemma 3. *Let $a, x > 0$ and $b \in \mathbb{R}$. If $x \leq a \log x + b$, then $x \leq 2a \log 2a - 2a + 2b$.*

Proof. Applying the basic inequality $\log u \leq u - 1$ with $u = x/(2a)$ gives

$$\log x \leq \frac{x}{2a} + \log 2a - 1,$$

which together with $x \leq a \log x + b$ implies that

$$x \leq \frac{x}{2} + a \log 2a - a + b.$$

We obtain the desired result after rearranging the terms. □

Lemma 4. *Let $a > 0$, $x \geq 1$ and $b \geq 0$. If $x \leq a\sqrt{\log x} + b$, then $x \leq 2a\sqrt{\log(a^2+1)} + 2b + 1$.*

Proof. Plugging $u = x/a^2$ into $\log u \leq u - 1$ gives

$$\log x \leq \frac{x}{a^2} + \log a^2 - 1 \leq \frac{x}{a^2} + \log(a^2 + 1),$$

which, combined with $x \leq a\sqrt{\log x} + b$ and $\sqrt{A+B} \leq \sqrt{A} + \sqrt{B}$ for any $A, B > 0$, implies

$$x \leq a\sqrt{\frac{x}{a^2} + \log(a^2+1)} + b \leq \sqrt{x} + a\sqrt{\log(a^2+1)} + b.$$

Solving the above inequality with Lemma 2 gives the desired result. □

Lemma 5. *Let $\{a_k\}_{k\geq 1}$ be a nonnegative sequence and define $A_k = A_0 + \sum_{i=1}^k a_i$ with $A_0 > 0$. Suppose there exists $C > 0$ such that $a_k \leq CA_{k-1}$ for all $k \geq 1$. Then the following inequalities hold for all $t \geq 1$:*

$$2(\sqrt{A_t} - \sqrt{A_0}) \leq \sum_{k=1}^t \frac{a_k}{\sqrt{A_{k-1}}} \leq (1 + \sqrt{1+C})(\sqrt{A_t} - \sqrt{A_0}), \tag{A.1}$$

$$\sum_{k=1}^t \frac{a_k}{A_{k-1}} \leq (1+C) \log \frac{A_t}{A_0}. \tag{A.2}$$

Proof. Firstly, we know from $A_0 > 0$ that the two summations in (A.1) and (A.2) are well-defined. Using the identity $a_k = A_k - A_{k-1} = (\sqrt{A_k} + \sqrt{A_{k-1}})(\sqrt{A_k} - \sqrt{A_{k-1}})$ and $\sqrt{A_k} \geq \sqrt{A_{k-1}}$, we obtain

$$\frac{a_k}{\sqrt{A_{k-1}}} = \Big(1 + \frac{\sqrt{A_k}}{\sqrt{A_{k-1}}}\Big)(\sqrt{A_k} - \sqrt{A_{k-1}}) \geq 2(\sqrt{A_k} - \sqrt{A_{k-1}}). \tag{A.3}$$

On the other hand, under the assumption $a_k \leq CA_{k-1}$, we have

$$\frac{\sqrt{A_k}}{\sqrt{A_{k-1}}} = \sqrt{1 + \frac{a_k}{A_{k-1}}} \leq \sqrt{1+C},$$

which implies that

$$\frac{a_k}{\sqrt{A_{k-1}}} \leq (1 + \sqrt{1+C})(\sqrt{A_k} - \sqrt{A_{k-1}}). \tag{A.4}$$

Summing (A.3) and (A.4) over $k = 1, \ldots, t$ gives (A.1).

We proceed to prove (A.2). Firstly, it follows from $a_k \leq CA_{k-1}$ that

$$A_k = A_{k-1} + a_k \leq (1+C)A_{k-1},$$

which implies that

$$\frac{a_k}{A_{k-1}} \leq (1+C)\frac{a_k}{A_k}. \tag{A.5}$$

Since the function $\log x$ is concave, we know that

$$\log A_k - \log A_{k-1} \geq \log'(A_k)(A_k - A_{k-1}) = \frac{a_k}{A_k}.$$

Applying the above inequality to (A.5) and summing over $k = 1, \ldots, t$ leads to

$$\begin{aligned}\sum_{k=1}^{t} \frac{a_k}{A_{k-1}} \leq & (1+C)\sum_{k=1}^{t}\frac{a_k}{A_k} \leq (1+C)\sum_{k=1}^{t}\big(\log A_k - \log A_{k-1}\big)\\ = & (1+C)(\log A_t - \log A_0) = (1+C)\log\frac{A_t}{A_0}.\end{aligned}$$

The proof is complete. □

B Proofs for Section 3

We first introduce the following classical descent lemma for PG-type updates, which serves as a starting point for our analysis in this section. Its proof can be found in the literature (see, e.g., [35]). We provide a proof here for completeness.

Lemma 6. *Suppose f is L-smooth. Let $\{x_k\}_{k\geq 1}$ be the sequence generated by the iterative scheme (PG). Then for any $k \geq 1$, it holds that*

$$F(x_{k+1}) \leq F(x_k) + \langle \nabla f(x_k) - g_k, x_{k+1} - x_k\rangle - \frac{1}{\eta_k}\|x_{k+1} - x_k\|^2 + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Proof. From the optimality condition of (PG), we have

$$0 \in g_k + \frac{1}{\eta_k}(x_{k+1} - x_k) + \partial h(x_{k+1}),$$

which, together with the convexity of h, implies that

$$h(x) \geq h(x_{k+1}) + \left\langle g_k + \frac{1}{\eta_k}(x_{k+1} - x_k), x_{k+1} - x\right\rangle, \ \forall x \in \mathbb{R}^d. \tag{B.1}$$

Setting $x = x_k$ in (B.1) and rearranging the terms gives

$$h(x_{k+1}) \leq h(x_k) - \langle g_k, x_{k+1} - x_k\rangle - \frac{1}{\eta_k}\|x_{k+1} - x_k\|^2. \tag{B.2}$$

On the other hand, it follows from the L-smoothness of f that

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{B.3}$$

Combining (B.2) and (B.3) leads to the desired result, which completes the proof. □

To prove Theorem 1, we also need the following lemma, which shows that the objective values at the iterates produced by Algorithm 1 are uniformly upper bounded. Recall that $\Delta_t := F(x_t) - F^*$.

Lemma 7. *Suppose f is L-smooth and Assumption 1 holds. Let $\{x_k\}_{k\geq 1}$ be the iterates generated by Algorithm 1 with full gradient $g_k := \nabla f(x_k)$ and any $\gamma, \eta > 0$. Then for any $t \geq 1$, we have*

$$\Delta_t \leq M := \Delta_1 + \left[L(\eta^2 + D^2)\log\frac{L\eta}{2\gamma} + \frac{LD^2}{2}\right]_+.$$

Proof. Firstly, since $g_k = \nabla f(x_k)$ and $\|x_{k+1} - x_k\| = \eta_k\|\mathcal{G}_k\|$, we obtain from Lemma 6 that

$$\begin{aligned}F(x_{k+1}) &\leq F(x_k) - \frac{1}{\eta_k}\|x_{k+1} - x_k\|^2 + \frac{L}{2}\|x_{k+1} - x_k\|^2\\ &= F(x_k) - \eta_k\|\mathcal{G}_k\|^2 + \frac{L\eta_k^2}{2}\|\mathcal{G}_k\|^2.\end{aligned} \tag{B.4}$$

To proceed, summing (B.4) over $k = 1, \ldots, t$ and subtracting F^* from both sides yields

$$\Delta_{t+1} \leq \Delta_1 + \sum_{k=1}^{t} \Big(\frac{L}{2} - \frac{1}{\eta_k}\Big)\eta_k^2 \|\mathcal{G}_k\|^2. \tag{B.5}$$

We consider two cases. Firstly, if $S_1 = \gamma > \frac{L\eta}{2}$, then it follows that

$$\frac{1}{\eta_k} = \frac{S_k}{\eta} \geq \frac{S_1}{\eta} > \frac{L}{2}, \ \forall k \geq 1,$$

which implies that $\Delta_{t+1} \leq \Delta_1 \leq M$ for all $t \geq 1$. Otherwise, if $\gamma \leq \frac{L\eta}{2}$, then $t^* := \max\{k \in [t] : L/2 \geq 1/\eta_k\}$ is well-defined and hence $S_{t^*} \leq \frac{L\eta}{2}$. For $k > t^*$, we know from the definition of t^* that $L/2 \leq 1/\eta_k$. Therefore, it follows from (B.5) that

$$\Delta_{t+1} \leq \Delta_1 + \sum_{k=1}^{t^*} \eta_k^2 \Big(\frac{L}{2} - \frac{1}{\eta_k}\Big)\|\mathcal{G}_k\|^2 \leq \Delta_1 + \frac{L}{2}\sum_{k=1}^{t^*} \eta_k^2 \|\mathcal{G}_k\|^2 = \Delta_1 + \frac{L\eta^2}{2}\sum_{k=1}^{t^*} \frac{\|\mathcal{G}_k\|^2}{S_k^2}. \tag{B.6}$$

From Assumption 1, we know that

$$\|\mathcal{G}_k\|^2 = \frac{\|x_{k+1} - x_k\|^2}{\eta_k^2} \leq \frac{D^2}{\eta_k^2} = \frac{D^2}{\eta^2} S_k^2, \ \forall k \geq 1.$$

Applying (A.2) in Lemma 5 with $a_k := \|\mathcal{G}_k\|^2$, $A_k := S_{k+1}^2$, and $C := D^2/\eta^2$, we derive from (B.6) that

$$\begin{aligned}
\Delta_{t+1} \leq & \Delta_1 + \frac{L\eta^2}{2}\Big(\sum_{k=1}^{t^*-1} \frac{\|\mathcal{G}_k\|^2}{S_k^2} + \frac{\|\mathcal{G}_{t^*}\|^2}{S_{t^*}^2}\Big) \\
\leq & \Delta_1 + \frac{L\eta^2}{2}\Big(1 + \frac{D^2}{\eta^2}\Big)\log\frac{S_{t^*}^2}{S_1^2} + \frac{LD^2}{2} \\
\leq & \Delta_1 + L(\eta^2 + D^2)\log\frac{L\eta}{2\gamma} + \frac{LD^2}{2},
\end{aligned}$$

where the last inequality follows from $S_{t^*} \leq \frac{L\eta}{2}$ and $S_1 = \gamma$. The proof is thus complete. □

In the stochastic setting, for notational convenience, we denote the stochastic noises as

$$\delta_{k,j} := \nabla f(x_k) - \nabla \ell(x_k; \xi_{k_j}), \quad \delta_k := \frac{1}{b_k}\sum_{j=1}^{b_k} \delta_{k,j} = \nabla f(x_k) - g_k.$$

Then it follows from the independence among the batch samples $\{\xi_{k_j}\}_{j=1}^{b_k}$ and Assumption 2 that

$$\mathbb{E}_k\Big[\Big\|\sum_{j=1}^{b_k}\delta_{k,j}\Big\|^2\Big] = \sum_{i,j=1}^{b_k}\mathbb{E}_k\big[\langle\delta_{k,i}, \delta_{k,j}\rangle\big] = \sum_{j=1}^{b_k}\mathbb{E}_k\big[\|\delta_{k,j}\|^2\big] \leq \sigma^2 b_k,$$

where to obtain the second equality, we used $\mathbb{E}_k\big[\langle\delta_{k,i}, \delta_{k,j}\rangle\big] = \big\langle\mathbb{E}_k\big[\delta_{k,i}\big], \mathbb{E}_k\big[\delta_{k,j}\big]\big\rangle = 0$ for $i \neq j$ due to the independence between $\delta_{k,i}$ and $\delta_{k,j}$. This further implies that

$$\mathbb{E}_k\big[\|\delta_k\|^2\big] = \mathbb{E}_k\Big[\Big\|\frac{1}{b_k}\sum_{j=1}^{b_k}\delta_{k,j}\Big\|^2\Big] = \frac{1}{b_k^2}\mathbb{E}_k\Big[\Big\|\sum_{j=1}^{b_k}\delta_{k,j}\Big\|^2\Big] \leq \frac{\sigma^2}{b_k}. \tag{B.7}$$

B.1 Proof of Theorem 1

As shown in Appendix E.1, the iterates of Algorithm 1 satisfy Assumption 1(a) with a constant D that depends only on Δ_1, L, γ, and η. Therefore, we use this D below and apply Lemma 7. We start by rearranging (B.4) to obtain

$$\|\mathcal{G}_k\|^2 \leq \frac{1}{\eta_k}(F(x_k) - F(x_{k+1})) + \frac{L}{2}\eta_k\|\mathcal{G}_k\|^2 = \frac{1}{\eta_k}(\Delta_k - \Delta_{k+1}) + \frac{L}{2}\eta_k\|\mathcal{G}_k\|^2. \tag{B.8}$$

Summing (B.8) over $k = 1, \dots, t$ leads to

$$\begin{aligned}\sum_{k=1}^{t} \|\mathcal{G}_k\|^2 &\le \frac{\Delta_1}{\eta_1} + \sum_{k=2}^{t} \Big(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\Big)\Delta_k - \frac{\Delta_{t+1}}{\eta_t} + \frac{L}{2}\sum_{k=1}^{t} \eta_k \|\mathcal{G}_k\|^2 \\ &\le \frac{M}{\eta_1} + M\sum_{k=2}^{t} \Big(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\Big) + \frac{L}{2}\sum_{k=1}^{t} \eta_k \|\mathcal{G}_k\|^2 \\ &= \frac{M}{\eta_t} + \frac{L}{2}\sum_{k=1}^{t} \eta_k \|\mathcal{G}_k\|^2 = \frac{M}{\eta} S_t + \frac{L\eta}{2}\sum_{k=1}^{t} \frac{\|\mathcal{G}_k\|^2}{S_k}, \end{aligned} \tag{B.9}$$

where we used $\Delta_k \le M$ and $\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \ge 0$ since η_k is non-increasing. Applying (A.1) in Lemma 5 with $a_k := \|\mathcal{G}_k\|^2$, $A_k := S_{k+1}^2$, and $C := D^2/\eta^2$, we further derive from (B.9) that

$$\begin{aligned}\sum_{k=1}^{t} \|\mathcal{G}_k\|^2 &\le \frac{M}{\eta}\sqrt{\gamma^2 + \sum_{k=1}^{t} \|\mathcal{G}_k\|^2} + \frac{L\eta}{2}\Big(1 + \sqrt{1 + D^2/\eta^2}\Big)\Bigg(\sqrt{\gamma^2 + \sum_{k=1}^{t} \|\mathcal{G}_k\|^2} - \gamma\Bigg) \\ &\le \frac{M\gamma}{\eta} + \frac{M}{\eta}\sqrt{\sum_{k=1}^{t} \|\mathcal{G}_k\|^2} + \frac{L\eta}{2}\big(2 + D/\eta\big)\sqrt{\sum_{k=1}^{t} \|\mathcal{G}_k\|^2} \\ &= \frac{M\gamma}{\eta} + \Big(\frac{M}{\eta} + L\eta + \frac{LD}{2}\Big)\sqrt{\sum_{k=1}^{t} \|\mathcal{G}_k\|^2}, \end{aligned} \tag{B.10}$$

where we used the basic inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ to obtain the second inequality. Solving (B.10) with Lemma 2 gives

$$\sum_{k=1}^{t} \|\mathcal{G}_k\|^2 \le \Big(\frac{M}{\eta} + L\eta + \frac{LD}{2}\Big)^2 + \frac{2M\gamma}{\eta},$$

which completes the proof after dividing by t.

B.2 Proof of Lemma 1

Firstly, note from the Cauchy–Schwarz inequality and the definition of $\widetilde{\mathcal{G}}_k$ (2.6) that

$$\sum_{k=1}^{t} \langle \delta_k, x_{k+1} - x_k \rangle \le \sum_{k=1}^{t} \|\delta_k\| \|x_{k+1} - x_k\| = \eta \sum_{k=1}^{t} \|\delta_k\| \frac{\|\widetilde{\mathcal{G}}_k\|}{S_k}.$$

Taking expectations on both sides of the above inequality and applying the Cauchy–Schwarz inequality for expectations $|\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$, we further obtain

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{t} \langle \delta_k, x_{k+1} - x_k \rangle\Big] &\le \eta \sum_{k=1}^{t} \mathbb{E}\Big[\|\delta_k\| \frac{\|\widetilde{\mathcal{G}}_k\|}{S_k}\Big] \le \eta \sum_{k=1}^{t} \sqrt{\mathbb{E}[\|\delta_k\|^2]\mathbb{E}\Big[\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}\Big]} \\ &\le \sigma\eta \sum_{k=1}^{t} \frac{1}{\sqrt{b_k}} \sqrt{\mathbb{E}\Big[\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}\Big]} \le \sigma\eta \sqrt{\sum_{k=1}^{t} \frac{1}{b_k}} \sqrt{\mathbb{E}\Big[\sum_{k=1}^{t} \frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}\Big]}, \end{aligned} \tag{B.11}$$

where we used (B.7) to obtain the third inequality, and the last one follows from the Cauchy–Schwarz inequality. Applying (A.2) in Lemma 5 with $a_k := \|\widetilde{\mathcal{G}}_k\|^2$, $A_k := S_{k+1}^2$, and $C := D^2/\eta^2$ to bound $\sum_{k=1}^{t} \frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}$, we derive from (B.11) that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{t} \langle \delta_k, x_{k+1} - x_k \rangle\Big] &\le \sigma\eta \sqrt{\sum_{k=1}^{t} \frac{1}{b_k}} \sqrt{\Big(1 + \frac{D^2}{\eta^2}\Big)\mathbb{E}\Big[\log \frac{S_{t+1}^2}{\gamma^2}\Big]} \\ &\le \sigma \sqrt{2(\eta^2 + D^2)\sum_{k=1}^{t} \frac{1}{b_k}} \sqrt{\log \mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]}, \end{aligned}$$

where the last inequality follows from Jensen's inequality applied to the concave function $\log(\cdot)$. The proof is now complete.

B.3 Proof of Theorem 2

We start by applying Lemma 6 and $\|x_{k+1} - x_k\| = \eta_k \|\widetilde{\mathcal{G}}_k\|$ to obtain

$$F(x_{k+1}) \leq F(x_k) + \langle \delta_k, x_{k+1} - x_k \rangle - \eta_k \|\widetilde{\mathcal{G}}_k\|^2 + \frac{L\eta_k^2}{2} \|\widetilde{\mathcal{G}}_k\|^2. \tag{B.12}$$

Subtracting F^* from both sides of (B.12) and rearranging the terms, we have

$$\eta_k \|\widetilde{\mathcal{G}}_k\|^2 \leq \Delta_k - \Delta_{k+1} + \frac{L\eta_k^2}{2} \|\widetilde{\mathcal{G}}_k\|^2 + \langle \delta_k, x_{k+1} - x_k \rangle. \tag{B.13}$$

Summing (B.13) over $k = 1, \ldots, t$ yields

$$\begin{aligned} \eta \sum_{k=1}^{t} \frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k} \leq & \Delta_1 - \Delta_{t+1} + \frac{L\eta^2}{2} \sum_{k=1}^{t} \frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} + \sum_{k=1}^{t} \langle \delta_k, x_{k+1} - x_k \rangle \\ \leq & \Delta_1 + \frac{L}{2}(\eta^2 + D^2) \log \frac{S_{t+1}^2}{\gamma^2} + \sum_{k=1}^{t} \langle \delta_k, x_{k+1} - x_k \rangle, \end{aligned} \tag{B.14}$$

where the second inequality follows by applying (A.2) in Lemma 5 with $a_k := \|\widetilde{\mathcal{G}}_k\|^2$, $A_k := S_{k+1}^2$, and $C := D^2/\eta^2$.

Next, we bound $\mathbb{E}[S_{t+1}]$. Applying (A.1) in Lemma 5 to the LHS of (B.14) gives

$$2\eta(S_{t+1} - \gamma) \leq \eta \sum_{k=1}^{t} \frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k} \leq \Delta_1 + L(\eta^2 + D^2) \log \frac{S_{t+1}}{\gamma} + \sum_{k=1}^{t} \langle \delta_k, x_{k+1} - x_k \rangle. \tag{B.15}$$

For brevity, let $\kappa_t := \frac{\sigma}{4\gamma\eta} \sqrt{2(\eta^2 + D^2) \sum_{k=1}^{t} 1/b_k}$. Taking expectation on both sides of (B.15) and applying Lemma 1, we derive after rearranging the terms that

$$\begin{aligned} \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] &\leq 1 + \frac{\Delta_1}{2\gamma\eta} + \frac{L}{2\gamma\eta}(\eta^2 + D^2) \mathbb{E}\left[\log \frac{S_{t+1}}{\gamma}\right] + 2\kappa_t \sqrt{\log \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]} \\ &\leq 1 + \frac{\Delta_1}{2\gamma\eta} + \frac{L}{2\gamma\eta}(\eta^2 + D^2) \log \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] + \kappa_t \left(1 + \log \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]\right) \\ &= 1 + \frac{\Delta_1}{2\gamma\eta} + \kappa_t + \left(\frac{L}{2\gamma\eta}(\eta^2 + D^2) + \kappa_t\right) \log \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right], \end{aligned} \tag{B.16}$$

where the second inequality follows from the concavity of $\log(\cdot)$ and the basic inequality $\sqrt{x} \leq \frac{1}{2}(1 + x)$. Solving (B.16) via Lemma 3, we conclude that

$$\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] \leq \left(\frac{L}{\gamma\eta}(\eta^2 + D^2) + 2\kappa_t\right) \log \left(\frac{L}{\gamma\eta}(\eta^2 + D^2) + 2\kappa_t\right) + 2 + \frac{\Delta_1}{\gamma\eta} - \frac{L}{\gamma\eta}(\eta^2 + D^2),$$

which leads to the following after omitting the constants $\{\gamma, \eta, D, L, \Delta_1\}$:

$$\mathbb{E}[S_{t+1}] \lesssim 1 + \left(1 + \sigma \sqrt{\sum_{k=1}^{t} \frac{1}{b_k}}\right) \log \left(1 + \sigma \sqrt{\sum_{k=1}^{t} \frac{1}{b_k}}\right). \tag{B.17}$$

From the nonexpansiveness of the proximal operator (2.2), we have

$$\|\mathcal{G}_k - \widetilde{\mathcal{G}}_k\|^2 = \frac{1}{\eta_k^2} \left\| \operatorname{prox}_{\eta_k h}(x_k - \eta_k g_k) - \operatorname{prox}_{\eta_k h}\left(x_k - \eta_k \nabla f(x_k)\right) \right\|^2 \leq \|g_k - \nabla f(x_k)\|^2 = \|\delta_k\|^2. \tag{B.18}$$

Therefore, we obtain from the basic inequalities $\|u+v\|^2 \le 2\|u\|^2+2\|v\|^2$ and $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, together with (B.18), that

$$\begin{aligned}\mathbb{E}\left[\sqrt{\sum_{k=1}^t \|\mathcal{G}_k\|^2}\right] \le& \mathbb{E}\left[\sqrt{2\sum_{k=1}^t \|\widetilde{\mathcal{G}}_k\|^2 + 2\sum_{k=1}^t \|\mathcal{G}_k - \widetilde{\mathcal{G}}_k\|^2}\right] \\ \le& \sqrt{2}\mathbb{E}\left[\sqrt{\sum_{k=1}^t \|\widetilde{\mathcal{G}}_k\|^2}\right] + \mathbb{E}\left[\sqrt{2\sum_{k=1}^t \|\delta_k\|^2}\right] \\ \le& \sqrt{2}\mathbb{E}[S_{t+1}] + \sqrt{2\sum_{k=1}^t \mathbb{E}[\|\delta_k\|^2]} \le \sqrt{2}\mathbb{E}[S_{t+1}] + \sigma\sqrt{2\sum_{k=1}^t \frac{1}{b_k}},\end{aligned} \tag{B.19}$$

where we used $S_k^2 = \gamma^2 + \sum_{j=1}^{k-1} \|\widetilde{\mathcal{G}}_j\|^2$ and the concavity of $\sqrt{\cdot}$ to derive the third inequality, and the last one follows from (B.7).

Squaring both sides of (B.19) and using $(a+b)^2 \le 2a^2+2b^2$ together with (B.17), we obtain

$$\begin{aligned}\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^t \|\mathcal{G}_k\|^2}\right]\right)^2 \le& 4\big(\mathbb{E}[S_{t+1}]\big)^2 + 4\sigma^2\sum_{k=1}^t \frac{1}{b_k} \\ \lesssim& \left(1+\sigma^2\sum_{k=1}^t \frac{1}{b_k}\right)\left(1+\log^2\left(1+\sigma\sqrt{\sum_{k=1}^t \frac{1}{b_k}}\right)\right).\end{aligned}$$

Then Theorem 2 follows by dividing both sides by t.

Finally, we establish the high-probability bound in Remark 2 through Markov's inequality $\mathbb{P}(X \ge a) \le \mathbb{E}[X]/a$ for any nonnegative random variable X and $a > 0$, which is equivalent to

$$\mathbb{P}\big(X^2 \le a^2\big) \ge 1 - \mathbb{E}[X]/a.$$

Given $\epsilon \in (0,1)$, setting $a := \mathbb{E}[X]/\epsilon$ in the above inequality gives

$$\mathbb{P}\big(X^2 \le (\mathbb{E}[X])^2/\epsilon^2\big) \ge 1-\epsilon. \tag{B.20}$$

Letting $X := \sqrt{\sum_{k=1}^t \|\mathcal{G}_k\|^2}$ in (B.20), we obtain from (B.19) that, with probability at least $1-\epsilon$,

$$\begin{aligned}\sum_{k=1}^t \|\mathcal{G}_k\|^2 \le& \frac{\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^t \|\mathcal{G}_k\|^2}\right]\right)^2}{\epsilon^2} \le \frac{4\big(\mathbb{E}\big[S_{t+1}\big]\big)^2 + 4\sigma^2\sum_{k=1}^t \frac{1}{b_k}}{\epsilon^2} \\ \lesssim& \frac{1}{\epsilon^2}\left[1+\log^2\left(1+\sigma\sqrt{\sum_{k=1}^t \frac{1}{b_k}}\right) + \sigma^2\sum_{k=1}^t \frac{1}{b_k}\left(1+\log^2\left(1+\sigma\sqrt{\sum_{k=1}^t \frac{1}{b_k}}\right)\right)\right],\end{aligned}$$

where we used (B.17) and the basic inequality $(a+b)^2 \le 2a^2+2b^2$ to obtain the last inequality.

C Proofs for Section 4

The following lemma provides descent properties for PG-type algorithms under the convexity of f, which serve as starting points for our analysis in this section. Recall that $\Delta_k := F(x_k) - F^*$.

Lemma 8. *Suppose f in (1.1) is convex. Let $\{x_k\}_{k\ge 1}$ be the sequence generated by the iterative scheme (PG). Then for any $k \ge 1$, we have the following results:*

(i). If f is G-Lipschitz continuous, then

$$\begin{aligned}\Delta_{k+1} \le& \frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big) - \frac{1}{2\eta_k}\|x_{k+1} - x_k\|^2 \\ &+ 2G\|x_k - x_{k+1}\| + \langle \delta_k, x_k - x^*\rangle + \langle \delta_k, x_{k+1} - x_k\rangle.\end{aligned} \tag{C.1}$$

(ii). If f is L-smooth, then

$$\begin{aligned}\Delta_{k+1} \leq & \frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big) + \frac{1}{2}\Big(L - \frac{1}{\eta_k}\Big)\|x_{k+1} - x_k\|^2 \\ & + \langle \delta_k, x_k - x^*\rangle + \langle \delta_k, x_{k+1} - x_k\rangle.\end{aligned} \tag{C.2}$$

Proof. We start by using the convexity of f to obtain

$$\begin{aligned}\Delta_k &= f(x_k) - f(x^*) + h(x_k) - h(x^*) \\ &\leq \langle \nabla f(x_k), x_k - x^*\rangle + h(x_k) - h(x^*) \\ &= \langle \nabla f(x_k), x_k - x_{k+1}\rangle + \langle g_k, x_{k+1} - x^*\rangle + \langle \delta_k, x_{k+1} - x^*\rangle + h(x_k) - h(x^*).\end{aligned} \tag{C.3}$$

We first prove the Lipschitz continuous case (i). To this end, we bound the first term on the RHS of (C.3) via the convexity of f and $\|\nabla f(x_{k+1})\| \leq G$ as follows

$$\begin{aligned}\langle \nabla f(x_k), x_k - x_{k+1}\rangle &= \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \langle \nabla f(x_k) - \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \\ &\leq f(x_k) - f(x_{k+1}) + \|\nabla f(x_k) - \nabla f(x_{k+1})\|\|x_k - x_{k+1}\| \\ &\leq f(x_k) - f(x_{k+1}) + 2G\|x_k - x_{k+1}\|.\end{aligned} \tag{C.4}$$

For the second term in (C.3), plugging $x = x^*$ into (B.1) and rearranging the terms yields

$$\begin{aligned}\langle g_k, x_{k+1} - x^*\rangle &\leq \frac{1}{\eta_k}\langle x_k - x_{k+1}, x_{k+1} - x^*\rangle + h(x^*) - h(x_{k+1}) \\ =& \frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|x_{k+1} - x_k\|^2\big) + h(x^*) - h(x_{k+1}),\end{aligned} \tag{C.5}$$

where we used the identity $2\langle a - b, b - c\rangle = \|a - c\|^2 - \|a - b\|^2 - \|b - c\|^2$ for the equality. To bound the third term in (C.3), we decompose

$$\langle \delta_k, x_{k+1} - x^*\rangle = \langle \delta_k, x_k - x^*\rangle + \langle \delta_k, x_{k+1} - x_k\rangle. \tag{C.6}$$

Applying (C.4), (C.5), and (C.6) to (C.3), we obtain

$$\begin{aligned}\Delta_k \leq & \frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big) - \frac{1}{2\eta_k}\|x_{k+1} - x_k\|^2 + 2G\|x_k - x_{k+1}\| \\ & + \langle \delta_k, x_k - x^*\rangle + \langle \delta_k, x_{k+1} - x_k\rangle + \Delta_k - \Delta_{k+1},\end{aligned}$$

which simplifies to (C.1).

We proceed to prove (ii) for L-smooth f. In this case, we can obtain a tighter bound for the first term on the RHS of (C.3) using

$$\langle \nabla f(x_k), x_k - x_{k+1}\rangle \leq f(x_k) - f(x_{k+1}) + \frac{L}{2}\|x_k - x_{k+1}\|^2. \tag{C.7}$$

Then (C.2) follows after plugging (C.7), (C.5), and (C.6) into (C.3). The proof is now complete. □

In both theorem proofs below, Assumption 1(b) implies $\|x_{k+1} - x_k\| \leq \|x_{k+1} - x^*\| + \|x_k - x^*\| \leq 2D$, so Lemma 1 applies with $2D$ in place of D.

C.1 Proof of Theorem 3

We first establish a bound for $\mathbb{E}[S_{t+1}]$. To this end, note from the G-Lipschitz continuity of f that

$$\begin{aligned}f(x_{k+1}) &\leq f(x_k) + G\|x_{k+1} - x_k\|, \\ -\langle \nabla f(x_k), x_{k+1} - x_k\rangle &\leq \|\nabla f(x_k)\|\|x_{k+1} - x_k\| \leq G\|x_{k+1} - x_k\|.\end{aligned}$$

Summing up the above two inequalities gives

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + 2G\|x_{k+1} - x_k\|. \tag{C.8}$$

Combining (C.8) and (B.1) with $x = x_k$, we obtain

$$\begin{aligned} &F(x_{k+1}) = f(x_{k+1}) + h(x_{k+1}) \\ \leq &f(x_k) + h(x_k) - \frac{1}{\eta_k}\|x_{k+1} - x_k\|^2 + 2G\|x_{k+1} - x_k\| + \langle \delta_k, x_{k+1} - x_k\rangle \\ = &F(x_k) - \eta_k\|\widetilde{\mathcal{G}}_k\|^2 + 2G\eta_k\|\widetilde{\mathcal{G}}_k\| + \langle \delta_k, x_{k+1} - x_k\rangle, \end{aligned} \tag{C.9}$$

where we used $\|x_{k+1} - x_k\| = \eta_k\|\widetilde{\mathcal{G}}_k\|$ to obtain the last equality. Rearranging the terms in (C.9) and noting $\eta_k = \eta/S_k$ gives

$$\eta\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k} \leq F(x_k) - F(x_{k+1}) + 2G\eta\frac{\|\widetilde{\mathcal{G}}_k\|}{S_k} + \langle \delta_k, x_{k+1} - x_k\rangle. \tag{C.10}$$

Summing (C.10) over $k = 1, \ldots, t$ and applying (A.1) in Lemma 5 yields

$$\begin{aligned} 2\eta(S_{t+1} - \gamma) &\leq \eta\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k} \\ &\leq F(x_1) - F(x_{t+1}) + 2G\eta\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|}{S_k} + \sum_{k=1}^{t}\langle \delta_k, x_{k+1} - x_k\rangle \\ &\leq \Delta_1 + 2G\eta\sqrt{t}\sqrt{\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}} + \sum_{k=1}^{t}\langle \delta_k, x_{k+1} - x_k\rangle, \end{aligned} \tag{C.11}$$

where the last inequality follows from $F(x_{t+1}) \geq F^*$ and the Cauchy–Schwarz inequality. Taking expectation on both sides of (C.11) and using the concavity of $\sqrt{\cdot}$, we obtain

$$\begin{aligned} \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] &\leq 1 + \frac{\Delta_1}{2\gamma\eta} + \frac{G\sqrt{t}}{\gamma}\mathbb{E}\left[\sqrt{\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}}\right] + \frac{1}{2\gamma\eta}\mathbb{E}\Big[\sum_{k=1}^{t}\langle \delta_k, x_{k+1} - x_k\rangle\Big] \\ &\leq 1 + \frac{\Delta_1}{2\gamma\eta} + \frac{G\sqrt{t}}{\gamma}\sqrt{\mathbb{E}\left[\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}\right]} + \frac{1}{2\gamma\eta}\mathbb{E}\Big[\sum_{k=1}^{t}\langle \delta_k, x_{k+1} - x_k\rangle\Big]. \end{aligned} \tag{C.12}$$

Applying (A.2) in Lemma 5 and Lemma 1 with $b_k = 1$ to (C.12) yields

$$\begin{aligned} \mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] &\leq 1 + \frac{\Delta_1}{2\gamma\eta} + \frac{G\sqrt{t}}{\gamma}\sqrt{\Big(1 + \frac{4D^2}{\eta^2}\Big)\mathbb{E}\left[\log\frac{S_{t+1}^2}{\gamma^2}\right]} + \frac{\sigma}{2\gamma\eta}\sqrt{2(\eta^2 + 4D^2)t}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]} \\ &\leq 1 + \frac{\Delta_1}{2\gamma\eta} + \frac{2G + \sigma}{2\gamma\eta}\sqrt{2(\eta^2 + 4D^2)t}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}, \end{aligned}$$

where the last inequality follows from the concavity of $\log(\cdot)$. Solving the above inequality via Lemma 4, we obtain

$$\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] \leq \frac{2G + \sigma}{\gamma\eta}\sqrt{2(\eta^2 + 4D^2)t}\sqrt{\log\left(\frac{(2G + \sigma)^2(\eta^2 + 4D^2)t}{2\gamma^2\eta^2} + 1\right)} + 3 + \frac{\Delta_1}{\gamma\eta},$$

which leads to the following after omitting the constants $\{\gamma, \eta, D, G, \Delta_1\}$:

$$\mathbb{E}[S_{t+1}] \lesssim 1 + (1 + \sigma)\sqrt{t\log\big(1 + (1 + \sigma^2)t\big)} \lesssim (1 + \sigma)\sqrt{t\log\big(1 + (1 + \sigma^2)t\big)}. \tag{C.13}$$

Next, we estimate $\mathbb{E}[\sum_{k=1}^{t}\Delta_{k+1}]$ by using the bound on $\mathbb{E}[S_{t+1}]$. Firstly, by omitting the negative term in (C.1), we obtain

$$\begin{aligned} \Delta_{k+1} \leq &\frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big) + 2G\eta\frac{\|\widetilde{\mathcal{G}}_k\|}{S_k} \\ &+ \langle \delta_k, x_k - x^*\rangle + \langle \delta_k, x_{k+1} - x_k\rangle. \end{aligned} \tag{C.14}$$

Summing (C.14) over $k = 1, \dots, t$ and noting that

$$\begin{aligned}
&\sum_{k=1}^{t} \frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big) \\
=&\frac{\|x_1 - x^*\|^2}{2\eta_1} + \sum_{k=2}^{t}\Big(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\Big)\|x_k - x^*\|^2 - \frac{\|x_{t+1} - x^*\|^2}{2\eta_t} \\
\le&\frac{D^2}{2\eta_1} + D^2\sum_{k=2}^{t}\Big(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\Big) = \frac{D^2}{2\eta_t},
\end{aligned} \tag{C.15}$$

we obtain

$$\sum_{k=1}^{t}\Delta_{k+1} \le \frac{D^2}{2\eta}S_t + 2G\eta\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|}{S_k} + \sum_{k=1}^{t}\langle\delta_k, x_k - x^*\rangle + \sum_{k=1}^{t}\langle\delta_k, x_{k+1} - x_k\rangle. \tag{C.16}$$

Taking expectation on both sides of (C.16) and noting that $\mathbb{E}_k[\langle\delta_k, x_k - x^*\rangle] = 0$ for all $k \ge 1$ due to the unbiasedness of g_k, we have

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{t}\Delta_{k+1}\Big] &\le \frac{D^2}{2\eta}\mathbb{E}[S_t] + 2G\eta\mathbb{E}\Big[\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|}{S_k}\Big] + \mathbb{E}\Big[\sum_{k=1}^{t}\langle\delta_k, x_{k+1} - x_k\rangle\Big] \\
&\le \frac{D^2}{2\eta}\mathbb{E}[S_{t+1}] + 2G\eta\sqrt{t}\sqrt{\mathbb{E}\Big[\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2}\Big]} + \mathbb{E}\Big[\sum_{k=1}^{t}\langle\delta_k, x_{k+1} - x_k\rangle\Big] \\
&\le \frac{D^2}{2\eta}\mathbb{E}[S_{t+1}] + 2G\eta\sqrt{t}\sqrt{\Big(1 + \frac{4D^2}{\eta^2}\Big)\mathbb{E}\Big[\log\frac{S_{t+1}^2}{\gamma^2}\Big]} \\
&\quad + \sigma\sqrt{2(\eta^2 + 4D^2)t}\sqrt{\log\mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]} \\
&\le \frac{D^2}{2\eta}\mathbb{E}[S_{t+1}] + (2G + \sigma)\sqrt{2(\eta^2 + 4D^2)t}\sqrt{\log\mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]},
\end{aligned} \tag{C.17}$$

where the second inequality follows from the Cauchy–Schwarz inequality and the concavity of $\sqrt{\cdot}$, the third one follows from (A.2) in Lemma 5 and Lemma 1 with $b_k \equiv 1$, and the last one follows from the concavity of $\log(\cdot)$. Finally, applying (C.13) to (C.17) and hiding fixed constants, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k=1}^{t}\Delta_{k+1}\Big] &\lesssim (1+\sigma)\sqrt{t\log\big(1 + (1+\sigma^2)t\big)} + (1+\sigma)\sqrt{t\log\Big(1 + (1+\sigma)\sqrt{t\log\big(1 + (1+\sigma^2)t\big)}\Big)} \\
&\lesssim (1+\sigma)\sqrt{t\log\big(1 + (1+\sigma^2)t\big)},
\end{aligned}$$

where the last inequality follows from $\log\big(1 + (1+\sigma^2)t\big) \le (1+\sigma^2)t \le (1+\sigma)^2 t$. This together with the convexity of F implies that

$$\mathbb{E}[F(\bar{x}_t) - F^*] \le \frac{1}{t}\mathbb{E}\Big[\sum_{k=1}^{t}\Delta_{k+1}\Big] \lesssim \frac{(1+\sigma)\sqrt{\log\big(1 + (1+\sigma^2)t\big)}}{\sqrt{t}}.$$

The proof is now complete.

C.2 Proof of Theorem 4

We start by summing (C.2) over $k = 1, \ldots, t$ and noting $\|x_{k+1} - x_k\| = \eta\|\widetilde{\mathcal{G}}_k\|/S_k$ to obtain

$$\begin{aligned}\sum_{k=1}^{t}\Delta_{k+1} \leq& \sum_{k=1}^{t}\frac{1}{2\eta_k}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\big) + \frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} \\ &+ \sum_{k=1}^{t}\langle\delta_k, x_k - x^*\rangle + \sum_{k=1}^{t}\langle\delta_k, x_{k+1} - x_k\rangle \\ \leq& \frac{D^2}{2\eta}S_t + \frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} + \sum_{k=1}^{t}\langle\delta_k, x_k - x^*\rangle + \sum_{k=1}^{t}\langle\delta_k, x_{k+1} - x_k\rangle,\end{aligned} \tag{C.18}$$

where the second inequality follows from (C.15).

To bound the second term on the RHS of (C.18), we consider two cases. Firstly, if $S_1 := \gamma > L\eta$, then it follows that

$$\frac{1}{\eta_k} = \frac{S_k}{\eta} \geq \frac{S_1}{\eta} > L, \ \forall k \geq 1,$$

which further implies

$$\frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} \leq 0.$$

Otherwise, if $\gamma \leq L\eta$, then $t^* := \max\{k \in [t] : L \geq 1/\eta_k\}$ is well-defined. Since $\eta_k = \eta/S_k$ is non-increasing, we have $L \geq S_k/\eta$ for all $1 \leq k \leq t^*$ and hence $S_{t^*} \leq L\eta$. For $k > t^*$, we know from the definition of t^* that $L \leq 1/\eta_k$. Therefore, by applying (A.2) in Lemma 5 and noting that $\|\widetilde{\mathcal{G}}_k\| \leq 2DS_k/\eta$ for all $k \geq 1$, we obtain

$$\frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} \leq \frac{L\eta^2}{2}\bigg(\sum_{k=1}^{t^*-1}\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} + \frac{\|\widetilde{\mathcal{G}}_{t^*}\|^2}{S_{t^*}^2}\bigg) \leq \frac{L\eta^2}{2}\Big(1 + \frac{4D^2}{\eta^2}\Big)\log\frac{L^2\eta^2}{\gamma^2} + 2LD^2.$$

Combining the above two cases, we have

$$\frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_k\|^2}{S_k^2} \leq \Big[L(\eta^2 + 4D^2)\log\frac{L\eta}{\gamma} + 2LD^2\Big]_+. \tag{C.19}$$

To proceed, we take expectation on both sides of (C.18) and apply $\mathbb{E}_k[\langle\delta_k, x_k - x^*\rangle] = 0$ for all $k \geq 1$, (C.19), and Lemma 1 with $b_k = 1$ to obtain

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{t}\Delta_{k+1}\Big] \leq& \frac{D^2}{2\eta}\mathbb{E}[S_t] + \Big[L(\eta^2 + 4D^2)\log\frac{L\eta}{\gamma} + 2LD^2\Big]_+ + \mathbb{E}\Big[\sum_{k=1}^{t}\langle\delta_k, x_{k+1} - x_k\rangle\Big] \\ \leq& \frac{D^2}{2\eta}\mathbb{E}[S_{t+1}] + \Big[L(\eta^2 + 4D^2)\log\frac{L\eta}{\gamma} + 2LD^2\Big]_+ \\ &+ \sigma\sqrt{2(\eta^2 + 4D^2)t}\sqrt{\log\mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]}.\end{aligned} \tag{C.20}$$

Moreover, we have established from the analysis in the nonconvex smooth case (cf. (B.17)) and $b_k = 1$ that

$$\mathbb{E}[S_{t+1}] \lesssim 1 + \big(1 + \sigma\sqrt{t}\big)\log\big(1 + \sigma\sqrt{t}\big). \tag{C.21}$$

Plugging (C.21) into (C.20) and using $\sqrt{x} \leq \frac{1}{2}(1 + x)$ yields

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{t}\Delta_{k+1}\Big] \lesssim& 1 + \big(1 + \sigma\sqrt{t}\big)\log\big(1 + \sigma\sqrt{t}\big) + \sigma\sqrt{t}\sqrt{\log\big(1 + \sigma\sqrt{t}\big)} \\ \leq& 1 + \big(1 + \sigma\sqrt{t}\big)\log\big(1 + \sigma\sqrt{t}\big) + \frac{\sigma\sqrt{t}}{2}\Big(1 + \log\big(1 + \sigma\sqrt{t}\big)\Big) \\ \lesssim& 1 + \big(1 + \sigma\sqrt{t}\big)\log\big(1 + \sigma\sqrt{t}\big),\end{aligned}$$

which together with the convexity of F implies that

$$\mathbb{E}[F(\bar{x}_t) - F^*] \le \frac{1}{t}\mathbb{E}\Big[\sum_{k=1}^{t}\Delta_{k+1}\Big] \lesssim \frac{1+\log\big(1+\sigma\sqrt{t}\big)}{t} + \frac{\sigma\log\big(1+\sigma\sqrt{t}\big)}{\sqrt{t}}.$$

The proof is now complete.

D Proofs for Section 5

With slight abuse of notation, we denote

$$\Delta_k := F(y_k) - F^*, \quad \eta_k := \frac{\eta}{S_k}, \quad \widetilde{\mathcal{G}}_{z,k} := \frac{z_k - z_{k+1}}{\eta_k}$$

for the analysis of Algorithm 2 throughout this section. Then it follows from the update rule of S_{k+1} that

$$S_k^2 = \gamma^2 + \sum_{j=1}^{k-1}\|\widetilde{\mathcal{G}}_{z,j}\|^2. \tag{D.1}$$

To prove the main results, we first introduce the following descent lemma for Algorithm 2 under the convexity of f.

Lemma 9. *Suppose f in (1.1) is convex. Let $\{x_k, y_k, z_k\}_{k\ge1}$ be the iterates generated by Algorithm 2. Then, for any $k \ge 1$ and $x \in \mathbb{R}^d$, we have the following results:*

(i). If f is G-Lipschitz continuous, then

$$\begin{aligned} F(y_{k+1}) \le& \big(1-\theta_k\big)F(y_k) + \theta_k F(x) + \theta_k\langle\delta_k, z_{k+1} - z_k\rangle + \theta_k\langle\delta_k, z_k - x\rangle \\ &+ \frac{\theta_k^2}{2\eta_k}\big(\|z_k - x\|^2 - \|z_{k+1} - x\|^2\big) - \frac{\theta_k^2}{2\eta_k}\|z_{k+1} - z_k\|^2 + 2G\theta_k\|z_{k+1} - z_k\|. \end{aligned} \tag{D.2}$$

(ii). If f is L-smooth, then

$$\begin{aligned} F(y_{k+1}) \le& \big(1-\theta_k\big)F(y_k) + \theta_k F(x) + \theta_k\langle\delta_k, z_{k+1} - z_k\rangle + \theta_k\langle\delta_k, z_k - x\rangle \\ &+ \frac{\theta_k^2}{2\eta_k}\big(\|z_k - x\|^2 - \|z_{k+1} - x\|^2\big) + \frac{\theta_k^2}{2}\Big(L - \frac{1}{\eta_k}\Big)\|z_{k+1} - z_k\|^2. \end{aligned} \tag{D.3}$$

Proof. We first prove (i). Since f is convex and G-Lipschitz continuous, arguing as in the derivation of (C.8) gives

$$f(y_{k+1}) \le f(x_k) + \langle\nabla f(x_k), y_{k+1} - x_k\rangle + 2G\|y_{k+1} - x_k\|. \tag{D.4}$$

To bound $f(x_k)$, we obtain from the convexity of f that

$$\begin{aligned} f(x_k) - \big[(1-\theta_k)f(y_k) + \theta_k f(x)\big] =& \theta_k\big(f(x_k) - f(x)\big) + \big(1-\theta_k\big)\big(f(x_k) - f(y_k)\big) \\ \le& \theta_k\langle\nabla f(x_k), x_k - x\rangle + \big(1-\theta_k\big)\langle\nabla f(x_k), x_k - y_k\rangle \\ =& \langle\nabla f(x_k), x_k - \theta_k x - \big(1-\theta_k\big)y_k\rangle. \end{aligned} \tag{D.5}$$

Applying (D.5) to (D.4) leads to

$$f(y_{k+1}) \le \big(1-\theta_k\big)f(y_k) + \theta_k f(x) + \langle\nabla f(x_k), y_{k+1} - \theta_k x - \big(1-\theta_k\big)y_k\rangle + 2G\|y_{k+1} - x_k\|. \tag{D.6}$$

To proceed, note from the update rules of x_k and (5.2) that

$$y_{k+1} = x_k + \theta_k\big(z_{k+1} - z_k\big) = (1-\theta_k)y_k + \theta_k z_k + \theta_k\big(z_{k+1} - z_k\big) = \theta_k z_{k+1} + \big(1-\theta_k\big)y_k, \tag{D.7}$$

which together with (D.6) implies that

$$\begin{aligned} f(y_{k+1}) \le& \big(1-\theta_k\big)f(y_k) + \theta_k f(x) + \theta_k\langle\nabla f(x_k), z_{k+1} - x\rangle + 2G\theta_k\|z_{k+1} - z_k\| \\ =& \big(1-\theta_k\big)f(y_k) + \theta_k f(x) + \theta_k\langle g_k, z_{k+1} - x\rangle + \theta_k\langle\delta_k, z_{k+1} - z_k\rangle \\ &+ \theta_k\langle\delta_k, z_k - x\rangle + 2G\theta_k\|z_{k+1} - z_k\|. \end{aligned} \tag{D.8}$$

Next, we bound the inner product $\theta_k\langle g_k, z_{k+1}-x\rangle$ on the RHS of (D.8). To this end, we first note from the convexity of h and (D.7) that

$$h(y_{k+1}) \le \theta_k h(z_{k+1}) + \big(1-\theta_k\big)h(y_k),$$

which can be rearranged as (noting $\alpha_k := 1/\theta_k$)

$$h(z_{k+1}) \ge \alpha_k h(y_{k+1}) - \alpha_k\big(1-\theta_k\big)h(y_k). \tag{D.9}$$

By the optimality condition of z_{k+1}, $\eta_k = \eta/S_k$, and (D.9), we have

$$\begin{aligned}
h(x) \ge& h(z_{k+1}) + \Big\langle g_k + \frac{\theta_k}{\eta_k}\big(z_{k+1}-z_k\big), z_{k+1}-x\Big\rangle \\
\ge& \alpha_k h(y_{k+1}) - \alpha_k\big(1-\theta_k\big)h(y_k) + \langle g_k, z_{k+1}-x\rangle + \frac{\theta_k}{\eta_k}\langle z_{k+1}-z_k, z_{k+1}-x\rangle \\
=& \alpha_k h(y_{k+1}) - \alpha_k\big(1-\theta_k\big)h(y_k) + \langle g_k, z_{k+1}-x\rangle \\
&+ \frac{\theta_k}{2\eta_k}\big(\|z_{k+1}-x\|^2 + \|z_{k+1}-z_k\|^2 - \|z_k-x\|^2\big),
\end{aligned}$$

which gives the following after multiplying by θ_k and rearranging the terms:

$$\begin{aligned}
\theta_k\langle g_k, z_{k+1}-x\rangle \le& \theta_k h(x) + \big(1-\theta_k\big)h(y_k) - h(y_{k+1}) \\
&+ \frac{\theta_k^2}{2\eta_k}(\|z_k-x\|^2 - \|z_{k+1}-x\|^2 - \|z_{k+1}-z_k\|^2).
\end{aligned} \tag{D.10}$$

Applying (D.10) to (D.8) and rearranging the terms using $F = f + h$ leads to (D.2).

We proceed to prove (ii), whose proof largely follows that of (i). The main difference is that when f is convex and L-smooth, instead of (D.4), we have the following refined bound:

$$f(y_{k+1}) \le f(x_k) + \langle \nabla f(x_k), y_{k+1}-x_k\rangle + \frac{L}{2}\|y_{k+1}-x_k\|^2.$$

Then, following the arguments from (D.5) to (D.8), we obtain

$$\begin{aligned}
f(y_{k+1}) \le& \big(1-\theta_k\big)f(y_k) + \theta_k f(x) + \theta_k\langle g_k, z_{k+1}-x\rangle + \theta_k\langle \delta_k, z_{k+1}-z_k\rangle \\
&+ \theta_k\langle \delta_k, z_k - x\rangle + \frac{L\theta_k^2}{2}\|z_{k+1}-z_k\|^2,
\end{aligned}$$

which together with (D.10) leads to (D.3). The proof is thus complete. □

To estimate the summation of inner products involving the stochastic noise, we also need the following lemma, whose proof is similar to that of Lemma 1.

Lemma 10. *Suppose Assumptions 2 and 3 hold. Let $\{x_k, y_k, z_k\}_{k\ge 1}$ be the iterates generated by Algorithm 2 with unit batch sizes $b_k \equiv 1$ and any $\gamma, \eta > 0$. Then for any $t \ge 1$, we have*

$$\mathbb{E}\Big[\sum_{k=1}^{t}\alpha_k\langle \delta_k, z_{k+1}-z_k\rangle\Big] \le \sigma\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]}.$$

Proof. Firstly, note from the Cauchy–Schwarz inequality and $\|z_{k+1}-z_k\| = \eta\|\widetilde{\mathcal{G}}_{z,k}\|/S_k$ that

$$\sum_{k=1}^{t}\alpha_k\langle \delta_k, z_{k+1}-z_k\rangle \le \sum_{k=1}^{t}\alpha_k\|\delta_k\|\|z_{k+1}-z_k\| = \eta\sum_{k=1}^{t}\alpha_k\|\delta_k\|\frac{\|\widetilde{\mathcal{G}}_{z,k}\|}{S_k}.$$

Taking expectations on both sides and applying the Cauchy–Schwarz inequality for expectations $|\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{t}\alpha_k\langle \delta_k, z_{k+1}-z_k\rangle\Big] \\
\le&\eta\sum_{k=1}^{t}\alpha_k\mathbb{E}\Big[\|\delta_k\|\frac{\|\widetilde{\mathcal{G}}_{z,k}\|}{S_k}\Big] \le \eta\sum_{k=1}^{t}\alpha_k\sqrt{\mathbb{E}[\|\delta_k\|^2]\mathbb{E}\Big[\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2}\Big]} \\
\le&\sigma\eta\sum_{k=1}^{t}\alpha_k\sqrt{\mathbb{E}\Big[\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2}\Big]} \le \sigma\eta\sqrt{\sum_{k=1}^{t}\alpha_k^2}\sqrt{\mathbb{E}\Big[\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2}\Big]},
\end{aligned} \tag{D.11}$$

where we used $\mathbb{E}_k[\|\delta_k\|^2] \le \sigma^2$ for the third inequality, and the Cauchy–Schwarz inequality for the last one. Applying (A.2) in Lemma 5 with $a_k := \|\widetilde{\mathcal{G}}_{z,k}\|^2$, $A_k := S_{k+1}^2$, and $C := 4D^2/\eta^2$, we further derive from (D.11) that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{t}\alpha_k\langle\delta_k, z_{k+1}-z_k\rangle\Big] \le& \sigma\eta\sqrt{\sum_{k=1}^{t}\alpha_k^2}\sqrt{\Big(1+\frac{4D^2}{\eta^2}\Big)\mathbb{E}\Big[\log\frac{S_{t+1}^2}{\gamma^2}\Big]}\\ \le& \sigma\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\Big[\frac{S_{t+1}}{\gamma}\Big]},\end{aligned}$$

where the last inequality follows from the concavity of $\log(\cdot)$. The proof is now complete. □

In our analysis, we leverage the following properties of the weights $\{\alpha_k\}_{k\ge1}$:

$$\alpha_k^2 - \alpha_k = \alpha_{k-1}^2 \ \textit{and}\ \frac{k+1}{2} \le \alpha_k \le k+1, \quad \forall k \ge 1. \tag{D.12}$$

Both results in (D.12) follow from the definition of α_k: the first identity is the quadratic equation in α_k, while the inequality can be verified by induction.

D.1 Proof of Theorem 5

We start by plugging $x = x^*$ into (D.2) to obtain

$$\begin{aligned}\Delta_{t+1} \le& (1-\theta_t)\Delta_t + \theta_t\langle\delta_t, z_{t+1}-z_t\rangle + \theta_t\langle\delta_t, z_t - x^*\rangle + \frac{\theta_t^2}{2\eta_t}\big(\|z_t-x^*\|^2 - \|z_{t+1}-x^*\|^2\big)\\ &- \frac{\theta_t^2}{2\eta_t}\|z_{t+1}-z_t\|^2 + 2G\theta_t\|z_{t+1}-z_t\|.\end{aligned} \tag{D.13}$$

Multiplying both sides of (D.13) by α_t^2 and applying $\alpha_t\theta_t = 1$ gives

$$\begin{aligned}\alpha_t^2\Delta_{t+1} \le& (\alpha_t^2-\alpha_t)\Delta_t + \alpha_t\langle\delta_t, z_{t+1}-z_t\rangle + \alpha_t\langle\delta_t, z_t-x^*\rangle\\ &+ \frac{1}{2\eta_t}\big(\|z_t-x^*\|^2 - \|z_{t+1}-x^*\|^2\big)\\ &- \frac{1}{2\eta_t}\|z_{t+1}-z_t\|^2 + 2G\alpha_t\|z_{t+1}-z_t\|,\end{aligned}$$

which together with the equality in (D.12) and $\|z_{t+1}-z_t\| = \eta_t\|\widetilde{\mathcal{G}}_{z,t}\| = \eta\|\widetilde{\mathcal{G}}_{z,t}\|/S_t$ yields

$$\begin{aligned}\alpha_t^2\Delta_{t+1} \le& \alpha_{t-1}^2\Delta_t + \alpha_t\langle\delta_t, z_{t+1}-z_t\rangle + \alpha_t\langle\delta_t, z_t-x^*\rangle + \frac{1}{2\eta_t}\big(\|z_t-x^*\|^2 - \|z_{t+1}-x^*\|^2\big)\\ &- \frac{\eta}{2}\frac{\|\widetilde{\mathcal{G}}_{z,t}\|^2}{S_t} + 2G\eta\alpha_t\frac{\|\widetilde{\mathcal{G}}_{z,t}\|}{S_t}.\end{aligned} \tag{D.14}$$

Applying (D.14) recursively and noting from Assumption 3 that

$$\begin{aligned}&\sum_{k=1}^{t}\frac{1}{2\eta_k}\big(\|z_k-x^*\|^2 - \|z_{k+1}-x^*\|^2\big)\\ =&\frac{\|z_1-x^*\|^2}{2\eta_1} + \sum_{k=2}^{t}\Big(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\Big)\|z_k-x^*\|^2 - \frac{\|z_{t+1}-x^*\|^2}{2\eta_t}\\ \le&\frac{D^2}{2\eta_1} + D^2\sum_{k=2}^{t}\Big(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\Big) = \frac{D^2}{2\eta_t} \le \frac{D^2}{2\eta}S_{t+1},\end{aligned} \tag{D.15}$$

we further obtain

$$\begin{aligned}\alpha_t^2\Delta_{t+1} \le& \alpha_0^2\Delta_1 - \frac{\eta}{2}\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k} + \frac{D^2}{2\eta}S_{t+1} + 2G\eta\sum_{k=1}^{t}\alpha_k\frac{\|\widetilde{\mathcal{G}}_{z,k}\|}{S_k}\\ &+ \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_{k+1}-z_k\rangle + \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_k-x^*\rangle.\end{aligned} \tag{D.16}$$

To proceed, from the definition of S_t (D.1), we apply (A.1) in Lemma 5 to derive

$$\frac{\eta}{2}\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k} \geq \eta\big(S_{t+1}-\gamma\big). \tag{D.17}$$

Additionally, since $\|\widetilde{\mathcal{G}}_{z,t}\|^2 = \|z_{t+1}-z_t\|^2/\eta_t^2 \leq 4D^2S_t^2/\eta^2$, it follows from the Cauchy–Schwarz inequality and (A.2) in Lemma 5 that

$$\begin{aligned}
&2G\eta\sum_{k=1}^{t}\alpha_k\frac{\|\widetilde{\mathcal{G}}_{z,k}\|}{S_k} \leq 2G\eta\sqrt{\sum_{k=1}^{t}\alpha_k^2}\sqrt{\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2}}\\
\leq&2G\eta\sqrt{\sum_{k=1}^{t}\alpha_k^2}\sqrt{\Big(1+\frac{4D^2}{\eta^2}\Big)\log\frac{S_{t+1}^2}{\gamma^2}} \leq 2G\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\frac{S_{t+1}}{\gamma}}.
\end{aligned} \tag{D.18}$$

Applying (D.17) and (D.18) to (D.16) leads to

$$\begin{aligned}
\alpha_t^2\Delta_{t+1} \leq&\alpha_0^2\Delta_1+\gamma\eta-\Big(\eta-\frac{D^2}{2\eta}\Big)S_{t+1}+2G\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\frac{S_{t+1}}{\gamma}}\\
&+\sum_{k=1}^{t}\alpha_k\langle\delta_k,z_{k+1}-z_k\rangle+\sum_{k=1}^{t}\alpha_k\langle\delta_k,z_k-x^*\rangle.
\end{aligned} \tag{D.19}$$

Taking expectations on both sides of (D.19) and noting that $\eta-\frac{D^2}{2\eta}>0$ (since $\eta>\frac{\sqrt{2}}{2}D$) and that $\mathbb{E}_k[\langle\delta_k,z_k-x^*\rangle]=0$ for all $k\geq 1$, we derive

$$\begin{aligned}
&\alpha_t^2\mathbb{E}[\Delta_{t+1}] \leq \alpha_0^2\Delta_1+\gamma\eta+2G\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\mathbb{E}\left[\sqrt{\log\frac{S_{t+1}}{\gamma}}\right]+\mathbb{E}\left[\sum_{k=1}^{t}\alpha_k\langle\delta_k,z_{k+1}-z_k\rangle\right]\\
\leq&\alpha_0^2\Delta_1+\gamma\eta+2G\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}+\sigma\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}\\
=&\alpha_0^2\Delta_1+\gamma\eta+(2G+\sigma)\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]},
\end{aligned} \tag{D.20}$$

where the second inequality follows from the concavity of $\sqrt{\log(\cdot)}$ and Lemma 10.

Next, we bound $\mathbb{E}[S_{t+1}]$. To this end, rearranging (D.19) gives

$$\begin{aligned}
\Big(\eta-\frac{D^2}{2\eta}\Big)S_{t+1} \leq&\alpha_0^2\Delta_1+\gamma\eta-\alpha_t^2\Delta_{t+1}+2G\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\frac{S_{t+1}}{\gamma}}\\
&+\sum_{k=1}^{t}\alpha_k\langle\delta_k,z_{k+1}-z_k\rangle+\sum_{k=1}^{t}\alpha_k\langle\delta_k,z_k-x^*\rangle.
\end{aligned} \tag{D.21}$$

Taking expectations on both sides of (D.21) and using $\Delta_{t+1}\geq 0$, the concavity of $\sqrt{\log(\cdot)}$, and Lemma 10, we obtain

$$\begin{aligned}
&\Big(\eta-\frac{D^2}{2\eta}\Big)\mathbb{E}[S_{t+1}]\\
\leq&\alpha_0^2\Delta_1+\gamma\eta+2G\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\mathbb{E}\left[\sqrt{\log\frac{S_{t+1}}{\gamma}}\right]+\mathbb{E}\left[\sum_{k=1}^{t}\alpha_k\langle\delta_k,z_{k+1}-z_k\rangle\right]\\
\leq&\alpha_0^2\Delta_1+\gamma\eta+(2G+\sigma)\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}.
\end{aligned} \tag{D.22}$$

Let $\rho := 1 - \frac{D^2}{2\eta^2} \in (0,1)$. Dividing both sides of (D.22) by $\gamma\eta\rho$ leads to

$$\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] \leq \frac{\alpha_0^2\Delta_1}{\gamma\eta\rho} + \frac{1}{\rho} + \frac{2G+\sigma}{\gamma\eta\rho}\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}.$$

Solving the above inequality using Lemma 4 yields

$$\begin{aligned}\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] \leq& \frac{4G+2\sigma}{\gamma\eta\rho}\sqrt{2(\eta^2+4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\left(\frac{2(2G+\sigma)^2(\eta^2+4D^2)}{\gamma^2\eta^2\rho^2}\sum_{k=1}^{t}\alpha_k^2+1\right)}\\ &+\frac{2\alpha_0^2\Delta_1}{\gamma\eta\rho}+\frac{2}{\rho}+1,\end{aligned}$$

which leads to the following after omitting the constants $\{\gamma, \eta, D, G, \alpha_0^2\Delta_1\}$ and noting $\alpha_k \leq k+1$ for all $k \geq 1$:

$$\mathbb{E}[S_{t+1}] \lesssim 1+(1+\sigma)\sqrt{\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\left(1+(1+\sigma^2)\sum_{k=1}^{t}\alpha_k^2\right)} \lesssim (1+\sigma)t^{3/2}\sqrt{\log\left(1+(1+\sigma^2)t^3\right)}. \tag{D.23}$$

Finally, applying (D.23) and $\alpha_k \leq k+1$ to (D.20) implies

$$\begin{aligned}\alpha_t^2\mathbb{E}[\Delta_{t+1}] &\lesssim 1+(1+\sigma)\sqrt{\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}\\ &\lesssim (1+\sigma)t^{3/2}\sqrt{\log\left(1+(1+\sigma)t^{3/2}\sqrt{\log\left(1+(1+\sigma^2)t^3\right)}\right)}\\ &\lesssim (1+\sigma)t^{3/2}\sqrt{\log\left(1+(1+\sigma^2)t^3\right)},\end{aligned} \tag{D.24}$$

where the last inequality follows from $\log\left(1+(1+\sigma^2)t^3\right) \leq (1+\sigma^2)t^3 \lesssim (1+\sigma)^2t^3$. Dividing both sides of (D.24) by α_t^2 and using $\alpha_t \geq \frac{t+1}{2}$ yields

$$\mathbb{E}[\Delta_{t+1}] \lesssim \frac{1+\sigma}{\sqrt{t}}\sqrt{\log\left(1+(1+\sigma^2)t^3\right)},$$

which completes the proof.

D.2 Proof of Theorem 6

We start by plugging $x = x^*$ into (D.3) to obtain

$$\begin{aligned}\Delta_{t+1} \leq& \big(1-\theta_t\big)\Delta_t + \theta_t\langle\delta_t, z_{t+1}-z_t\rangle + \theta_t\langle\delta_t, z_t - x^*\rangle\\ &+\frac{\theta_t^2}{2\eta_t}\big(\|z_t-x^*\|^2 - \|z_{t+1}-x^*\|^2\big) + \frac{\theta_t^2}{2}\Big(L-\frac{1}{\eta_t}\Big)\|z_{t+1}-z_t\|^2.\end{aligned}$$

Multiplying both sides of the above inequality by α_t^2 and using the equality in (D.12) leads to

$$\begin{aligned}\alpha_t^2\Delta_{t+1} \leq& \alpha_{t-1}^2\Delta_t + \frac{1}{2}\Big(L-\frac{1}{\eta_t}\Big)\|z_{t+1}-z_t\|^2 + \frac{1}{2\eta_t}\big(\|z_t-x^*\|^2 - \|z_{t+1}-x^*\|^2\big)\\ &+\alpha_t\langle\delta_t, z_{t+1}-z_t\rangle + \alpha_t\langle\delta_t, z_t-x^*\rangle,\end{aligned}$$

and applying this inequality recursively together with $\|z_{t+1}-z_t\| = \eta_t\|\widetilde{\mathcal{G}}_{z,t}\| = \eta\|\widetilde{\mathcal{G}}_{z,t}\|/S_t$ and (D.15) yields

$$\begin{aligned}\alpha_t^2\Delta_{t+1} \leq& \alpha_0^2\Delta_1 + \frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L-\frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2} + \frac{D^2}{2\eta}S_{t+1}\\ &+\sum_{k=1}^{t}\alpha_k\langle\delta_k, z_{k+1}-z_k\rangle + \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_k-x^*\rangle.\end{aligned} \tag{D.25}$$

To bound the first summation term on the RHS of (D.25), we consider two cases. Firstly, if $S_1 := \gamma > L\eta$, then it follows that $1/\eta_k = S_k/\eta \geq S_1/\eta > L$ for all $k \geq 1$, which implies that

$$\frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2} \leq 0.$$

Otherwise, if $\gamma \leq L\eta$, then $t^* := \max\{k \in [t] : L \geq 1/\eta_k\}$ is well-defined. Since $\eta_k = \eta/S_k$ is non-increasing, we have $L \geq S_k/\eta$ for all $1 \leq k \leq t^*$ and hence $S_{t^*} \leq L\eta$. For $k > t^*$, we know from the definition of t^* that $L \leq 1/\eta_k$. Therefore, by applying (A.2) in Lemma 5 and noting that $\|\widetilde{\mathcal{G}}_{z,k}\|^2 = \|z_{k+1} - z_k\|^2/\eta_k^2 \leq 4D^2S_k^2/\eta^2$ for all $k \geq 1$, we obtain

$$\begin{aligned}
&\frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2} \leq \frac{L\eta^2}{2}\Big(\sum_{k=1}^{t^*-1}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2} + \frac{\|\widetilde{\mathcal{G}}_{z,t^*}\|^2}{S_{t^*}^2}\Big)\\
\leq&\frac{L\eta^2}{2}\bigg[\Big(1 + \frac{4D^2}{\eta^2}\Big)\log\frac{S_{t^*}^2}{\gamma^2} + \frac{4D^2}{\eta^2}\bigg] \leq L\big(\eta^2 + 4D^2\big)\log\frac{L\eta}{\gamma} + 2LD^2.
\end{aligned}$$

Combining the above two cases, we have

$$\frac{\eta^2}{2}\sum_{k=1}^{t}\Big(L - \frac{1}{\eta_k}\Big)\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2} \leq \Big[L(\eta^2 + 4D^2)\log\frac{L\eta}{\gamma} + 2LD^2\Big]_+. \tag{D.26}$$

Taking expectations on both sides of (D.25) and applying $\mathbb{E}_k[\langle\delta_k, z_k - x^*\rangle] = 0$ for all $k \geq 1$, Lemma 10, and the bound (D.26), we obtain

$$\begin{aligned}
\alpha_t^2\mathbb{E}[\Delta_{t+1}] \leq&\alpha_0^2\Delta_1 + \Big[L(\eta^2 + 4D^2)\log\frac{L\eta}{\gamma} + 2LD^2\Big]_+ + \frac{D^2\gamma}{2\eta}\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg]\\
&+ \sigma\sqrt{2(\eta^2 + 4D^2)\sum_{k=1}^{t}\alpha_k^2}\sqrt{\log\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg]}.
\end{aligned} \tag{D.27}$$

To proceed, we bound $\mathbb{E}\big[\frac{S_{t+1}}{\gamma}\big]$. To this end, we rearrange (D.25) and apply Lemma 5 to derive

$$\begin{aligned}
&\Big(\eta - \frac{D^2}{2\eta}\Big)S_{t+1} - \gamma\eta = \eta\big(S_{t+1} - \gamma\big) - \frac{D^2}{2\eta}S_{t+1} \leq \frac{\eta}{2}\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k} - \frac{D^2}{2\eta}S_{t+1}\\
\leq&\alpha_0^2\Delta_1 - \alpha_t^2\Delta_{t+1} + \frac{L\eta^2}{2}\sum_{k=1}^{t}\frac{\|\widetilde{\mathcal{G}}_{z,k}\|^2}{S_k^2} + \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_{k+1} - z_k\rangle + \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_k - x^*\rangle\\
\leq&\alpha_0^2\Delta_1 + \frac{L\eta^2}{2}\Big(1 + \frac{4D^2}{\eta^2}\Big)\log\frac{S_{t+1}^2}{\gamma^2} + \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_{k+1} - z_k\rangle + \sum_{k=1}^{t}\alpha_k\langle\delta_k, z_k - x^*\rangle
\end{aligned}$$

Since $\eta > \sqrt{2}D/2$, it follows that $\rho := 1 - \frac{D^2}{2\eta^2} \in (0, 1)$. For brevity, in this proof we define $\kappa_t := \frac{\sigma}{2\gamma\eta\rho}\sqrt{2(\eta^2 + 4D^2)\sum_{k=1}^{t}\alpha_k^2}$. Taking expectations on both sides of the above inequality and dividing by $\gamma\eta\rho$ yields

$$\begin{aligned}
\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg] \leq&\frac{1}{\rho} + \frac{\alpha_0^2\Delta_1}{\gamma\eta\rho} + \frac{L(\eta^2 + 4D^2)}{\gamma\eta\rho}\mathbb{E}\bigg[\log\frac{S_{t+1}}{\gamma}\bigg] + 2\kappa_t\sqrt{\log\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg]}\\
\leq&\frac{1}{\rho} + \frac{\alpha_0^2\Delta_1}{\gamma\eta\rho} + \frac{L(\eta^2 + 4D^2)}{\gamma\eta\rho}\log\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg] + \kappa_t\bigg(1 + \log\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg]\bigg)\\
=&\frac{1}{\rho} + \frac{\alpha_0^2\Delta_1}{\gamma\eta\rho} + \kappa_t + \bigg(\frac{L(\eta^2 + 4D^2)}{\gamma\eta\rho} + \kappa_t\bigg)\log\mathbb{E}\bigg[\frac{S_{t+1}}{\gamma}\bigg],
\end{aligned}$$

where the second inequality follows from the concavity of $\log(\cdot)$ and the basic inequality $\sqrt{x} \le \frac{1}{2}(1+x)$. Solving the above inequality using Lemma 3 and noting $\alpha_k \le k+1$ for all $k \ge 1$ gives

$$\begin{aligned}\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right] &\le \left(\frac{2L\left(\eta^2+4D^2\right)}{\gamma\eta\rho}+2\kappa_t\right)\log\left(\frac{2L(\eta^2+4D^2)}{\gamma\eta\rho}+2\kappa_t\right)\\ &\quad+\frac{2}{\rho}+\frac{2\alpha_0^2\Delta_1}{\gamma\eta\rho}-\frac{2L\left(\eta^2+4D^2\right)}{\gamma\eta\rho}\\ &\lesssim 1+\left(1+\sigma\sqrt{\sum_{k=1}^t\alpha_k^2}\right)\log\left(1+\sigma\sqrt{\sum_{k=1}^t\alpha_k^2}\right)\\ &\lesssim 1+\left(1+\sigma t^{3/2}\right)\log\left(1+\sigma t^{3/2}\right). \end{aligned} \tag{D.28}$$

Finally, applying (D.28) to (D.27) yields

$$\begin{aligned}\alpha_t^2\mathbb{E}[\Delta_{t+1}] &\lesssim 1+\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]+\sigma\sqrt{\sum_{k=1}^t\alpha_k^2}\sqrt{\log\mathbb{E}\left[\frac{S_{t+1}}{\gamma}\right]}\\ &\lesssim 1+\left(1+\sigma t^{3/2}\right)\log\left(1+\sigma t^{3/2}\right)+\sigma t^{3/2}\sqrt{\log\left(1+\left(1+\sigma t^{3/2}\right)\log\left(1+\sigma t^{3/2}\right)\right)}\\ &\lesssim 1+\log\left(1+\sigma t^{3/2}\right)+\sigma t^{3/2}\log\left(1+\sigma t^{3/2}\right).\end{aligned} \tag{D.29}$$

Dividing both sides of (D.29) by α_t^2 and using $\alpha_t \ge \frac{t+1}{2}$ yields

$$\mathbb{E}[\Delta_{t+1}] \lesssim \frac{1+\log\left(1+\sigma t^{3/2}\right)}{t^2}+\frac{\sigma\log\left(1+\sigma t^{3/2}\right)}{\sqrt{t}},$$

which completes the proof.

E Discussion

E.1 The boundedness assumption

Assumption 1 consists of two parts: *bounded steps* ($\|x_{k+1}-x_k\| \le D$), which is used in all settings, and *bounded iterates* ($\|x_k - x^*\| \le D$), which is used only when f is convex. We now show that the bounded-step part holds automatically for any smooth f in the deterministic setting, and hence can be removed from Theorem 1 and Corollary 2.

To this end, define the inverse step size $D_k := 1/\eta_k = S_k/\eta$, so that $\mathcal{G}_k = D_k(x_k - x_{k+1})$ and, from the update rule (2.5),

$$D_{k+1}^2 = D_k^2 + \frac{\|\mathcal{G}_k\|^2}{\eta^2}. \tag{E.1}$$

Plugging $g_k = \nabla f(x_k)$ into the descent inequality of Lemma 6 yields

$$F(x_{k+1}) \le F(x_k) - \left(D_k - \frac{L}{2}\right)\|x_{k+1}-x_k\|^2.$$

Summing over $k = 1, \dots, t$ and rearranging the terms, we obtain the telescoping inequality

$$\frac{1}{2}\sum_{k=1}^t D_k\|x_{k+1}-x_k\|^2 \le \Delta_1 + \frac{1}{2}\sum_{k=1}^t (L-D_k)\|x_{k+1}-x_k\|^2. \tag{E.2}$$

Since $\{D_k\}_{k\ge1}$ is nondecreasing, we consider two cases. If $D_k \le L$ for all $k \ge 1$, then it follows from (E.1) that

$$\sum_{k=1}^t \|\mathcal{G}_k\|^2 = \eta^2(D_{t+1}^2 - D_1^2) \le \eta^2(L^2 - D_1^2),$$

and hence

$$\sum_{k=1}^{t} \|x_{k+1} - x_k\|^2 = \sum_{k=1}^{t} \frac{\|\mathcal{G}_k\|^2}{D_k^2} \le \frac{\eta^2(L^2 - D_1^2)}{D_1^2} < \infty.$$

Otherwise, there exists $N \ge 0$ such that $D_k > L$ for all $k > N$, so $(L - D_k) \le 0$ for $k > N$ and the RHS of (E.2) is bounded by $\Delta_1 + \frac{L}{2}\sum_{k=1}^{N} \|x_{k+1} - x_k\|^2 < \infty$. In both cases, (E.2) together with $D_k \ge D_1$ implies that $\|x_{k+1} - x_k\|^2$ is uniformly bounded for all $k \ge 1$ by a constant depending only on Δ_1, L, γ, and η.

The bounded-iterate part holds if $\mathrm{dom}(h)$ is bounded or the operator $\arg\min_{x\in\mathbb{R}^d}\{\langle u, x\rangle + h(x)\}$ is bounded for all $u \in \mathbb{R}^d$ [12, Lemma 2]. For general composite problems, removing boundedness is challenging because of the delayed accumulation and the proximal structure. The sum inequality (A.2) requires $\|\widetilde{\mathcal{G}}_k\|^2 \le CS_k^2$, which bounded steps ensure through $\|\widetilde{\mathcal{G}}_k\|^2/S_k^2 = \|x_{k+1} - x_k\|^2/\eta^2$. Moreover, the descent term is $\langle \nabla f(x_k), x_{k+1} - x_k\rangle = -\eta_k\langle \nabla f(x_k), \widetilde{\mathcal{G}}_k\rangle$, which generally differs from $-\eta_k\langle \nabla f(x_k), g_k\rangle$. Thus, arguments based on the gradient-step identity in the non-composite setting do not directly extend to the proximal update.

A natural fix is to use the non-delayed accumulator

$$S_k^2 = \gamma^2 + \sum_{j=1}^{k} \|\widetilde{\mathcal{G}}_j\|^2.$$

This would avoid the bounded-step condition in the corresponding summation estimate, but $\widetilde{\mathcal{G}}_k$ depends on x_{k+1} and hence is unavailable until the proximal step is evaluated. An intuitive alternative is the hybrid scheme

$$S_k^2 = \gamma^2 + \|g_k\|^2 + \sum_{j=1}^{k-1} \|\widetilde{\mathcal{G}}_j\|^2,$$

which recovers non-delayed AdaGrad-Norm when $h = 0$. For general h, however, the gradient norm does not control the gradient mapping. Consequently, adding $\|g_k\|^2$ may contribute only a bounded offset: for example, for $f(x) = x^2/2$ and h the indicator function of the closed interval $[1, 2]$, we have $1 \le \|g_k\|^2 \le 4$ in the deterministic setting.

E.2 Block-diagonal extension

The scalar step size in Algorithm 1 can be extended to a block-diagonal (in particular, coordinate-wise) scheme. Partition $x \in \mathbb{R}^{md}$ into m blocks $x = (x_1, \dots, x_m)$ with $x_i \in \mathbb{R}^d$, and define $D_k = \mathrm{diag}(D_{k,1}I_d, \dots, D_{k,m}I_d)$ with $D_{k,i} > 0$ for $i \in [m]$. Replace the core update of Algorithm 1 with

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^{md}} \Big\{ \langle g_k, x\rangle + \frac{1}{2}\|x - x_k\|_{D_k}^2 + h(x) \Big\}, \tag{E.3}$$

where $\|x\|_{D_k}^2 := \sum_{i=1}^{m} D_{k,i}\|x_i\|^2$, and replace the scalar update of S_k with the block-wise update

$$D_{k+1,i}^2 = D_{k,i}^2\Big(1 + \frac{\|x_{k+1,i} - x_{k,i}\|^2}{\eta_i^2}\Big), \qquad i = 1, \dots, m. \tag{E.4}$$

Setting $d = 1$ and $m = d_{\mathrm{orig}}$ recovers the fully coordinate-wise AdaGrad scaling, while setting $d = d_{\mathrm{orig}}$ and $m = 1$ recovers Algorithm 1.

For the nonconvex smooth analysis, the gradient mapping of block i is $\mathcal{G}_{k,i} := D_{k,i}(x_{k,i} - x_{k+1,i})$, so $\|\mathcal{G}_{k,i}\|^2 = D_{k,i}^2\|x_{k+1,i} - x_{k,i}\|^2$ and $\|\mathcal{G}_k\|^2 = \sum_{i=1}^{m} \|\mathcal{G}_{k,i}\|^2$. Suppose f is $\mathcal{B}$-smooth in the sense that

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{1}{2}\|x - y\|_{\mathcal{B}}^2, \quad \forall x, y \in \mathbb{R}^{md},$$

with $\mathcal{B} = \mathrm{diag}(\beta_1 I_d, \dots, \beta_m I_d)$ (the scalar L-smoothness corresponds to $m = 1$ and $\beta_1 = L$). Then the descent inequality in Lemma 6 generalizes to

$$F(x_{k+1}) \le F(x_k) - \|x_{k+1} - x_k\|_{D_k}^2 + \frac{1}{2}\|x_{k+1} - x_k\|_{\mathcal{B}}^2,$$

which, after summing over $k = 1, \dots, t$ and rearranging, gives

$$\frac{1}{2}\sum_{k=1}^{t}\|x_{k+1}-x_k\|_{D_k}^2 \le \Delta_1 + \frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{t}(\beta_i - D_{k,i})\|x_{k+1,i}-x_{k,i}\|^2. \tag{E.5}$$

We bound the RHS of (E.5) by a per-block case split. Under boundedness $\|x_{k+1}-x_k\| \le R$, for each $i \in [m]$: if $D_{1,i} \ge \beta_i$, every term $(\beta_i - D_{k,i}) \le 0$ and contributes nothing; otherwise, let $t_i := \max\{k \in [t] : D_{k,i} \le \beta_i\}$. Applying the log-bound (A.2) in Lemma 5 with $a_k = \|\mathcal{G}_{k,i}\|^2/\eta_i^2$, $A_k = D_{k+1,i}^2$, and $C = R^2/\eta_i^2$ yields

$$\sum_{k=1}^{t_i-1}\|x_{k+1,i}-x_{k,i}\|^2 = \eta_i^2\sum_{k=1}^{t_i-1}\frac{\|\mathcal{G}_{k,i}\|^2/\eta_i^2}{D_{k,i}^2} \le 2(\eta_i^2+R^2)\log\frac{\beta_i}{D_{1,i}}.$$

Therefore, $\sum_{k=1}^{t}(\beta_i - D_{k,i})\|x_{k+1,i}-x_{k,i}\|^2 \le \beta_i\sum_{k=1}^{t_i}\|x_{k+1,i}-x_{k,i}\|^2 \le 2\beta_i(\eta_i^2+R^2)\log\frac{\beta_i}{D_{1,i}} + \beta_i R^2$, and combining both cases yields

$$\frac{1}{2}\sum_{k=1}^{t}\|x_{k+1}-x_k\|_{D_k}^2 \le \Delta_1 + \sum_{i=1}^{m}\left[\beta_i(\eta_i^2+R^2)\log\frac{\beta_i}{D_{1,i}} + \frac{\beta_i}{2}R^2\right]_+ := M_1. \tag{E.6}$$

On the other hand, applying the lower bound (A.1) in Lemma 5 with $a_k = \|\mathcal{G}_{k,i}\|^2/\eta_i^2$ and $A_k = D_{k+1,i}^2$ gives

$$\sum_{k=1}^{t}\|x_{k+1}-x_k\|_{D_k}^2 = \sum_{i=1}^{m}\eta_i^2\sum_{k=1}^{t}\frac{\|\mathcal{G}_{k,i}\|^2/\eta_i^2}{D_{k,i}} \ge 2\sum_{i=1}^{m}\eta_i^2(D_{t+1,i}-D_{1,i}),$$

which together with (E.6) implies (denoting $\underline{\eta}^2 := \min_{i\in[m]}\eta_i^2$)

$$D_{t+1,i} \le \sum_{i=1}^{m} D_{t+1,i} \le \frac{1}{\underline{\eta}^2}\sum_{i=1}^{m}\eta_i^2 D_{1,i} + \frac{M_1}{\underline{\eta}^2} := M_2, \quad \forall i \in [m].$$

Since $D_{k+1,i}^2 = D_{1,i}^2 + \frac{1}{\eta_i^2}\sum_{j=1}^{k}\|\mathcal{G}_{j,i}\|^2$, summing over $i = 1, \dots, m$ gives $\sum_{k=1}^{t}\|\mathcal{G}_k\|^2 = \sum_{i=1}^{m}\eta_i^2(D_{t+1,i}^2 - D_{1,i}^2)$, which together with the boundedness of $D_{t+1,i}$ implies that

$$\sum_{k=1}^{t}\|\mathcal{G}_k\|^2 \le \sum_{i=1}^{m}\eta_i^2(M_2+D_{1,i})(D_{t+1,i}-D_{1,i}) \le (M_2 + \max_i D_{1,i})M_1,$$

hence $\frac{1}{t}\sum_{k=1}^{t}\|\mathcal{G}_k\|^2 = O(1/t)$. The block-diagonal scheme extends to the convex settings with an analogous analysis, and the bounded-step derivation in Appendix E.1 carries over verbatim with β_i in place of L.

F Extension to Hölder Smoothness for Convex Functions

For convex objectives, Orabona [34] showed that AdaGrad can adapt to Hölder smoothness, which interpolates between smoothness and Lipschitzness. This universal adaptation was later extended in Rodomanov et al. [38]. Inspired by their work, we extend the convex analysis for both Algorithm 1 and Algorithm 2 under the Hölder-smooth condition.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is (L_ν, ν)-Hölder smooth with $L_\nu > 0$ and $\nu \in [0, 1]$ if

$$\|\nabla f(x) - \nabla f(y)\| \le L_\nu\|x-y\|^\nu, \qquad x, y \in \mathbb{R}^d.$$

The endpoint $\nu = 1$ gives standard L-smoothness; $\nu = 0$ includes differentiable G-Lipschitz functions with $L_0 = 2G$. Note that the Hölder smoothness condition also implies [34, Theorem 1]:

$$f(y) \le f(x) + \langle\nabla f(x), y-x\rangle + \frac{L_\nu}{1+\nu}\|y-x\|^{1+\nu}, \qquad x, y \in \mathbb{R}^d. \tag{F.1}$$

F.1 Convergence of Algorithm 1 under Hölder smoothness

Theorem 7. *Suppose f is convex and (L_ν, ν)-Hölder smooth for some $\nu \in [0,1]$, and Assumptions 1(b) and 2 hold. Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $b_k \equiv 1$ and any $\gamma, \eta > 0$. Then, for every $t \geq 1$, the averaged iterate $\bar{x}_t = t^{-1}\sum_{k=1}^{t} x_{k+1}$ satisfies*

$$\mathbb{E}[F(\bar{x}_t) - F^*] \lesssim \left(t^{-\frac{1+\nu}{2}} + \frac{\sigma}{\sqrt{t}}\right)\left[1 + \log\left(1 + t^{\frac{1-\nu}{2}} + \sigma\sqrt{t}\right)\right]^{\frac{1+\nu}{2}}.$$

Proof. For the cases $\nu = 0$ and $\nu = 1$, the result follows directly from Theorem 3 and Theorem 4, respectively. Hereafter, we assume $0 < \nu < 1$.

Recall that $\Delta_k := F(x_k) - F^*$ and $\delta_k := \nabla f(x_k) - g_k$, and let $d_k := \|x_{k+1} - x_k\|$. Since $S_{k+1}^2 = S_k^2(1 + d_k^2/\eta^2)$ and $d_k \leq 2D$ by Assumption 1(b), applying (A.2) in Lemma 5 with $a_k = S_k^2 d_k^2/\eta^2$, $A_k = S_{k+1}^2$, and $C = 4D^2/\eta^2$ then gives

$$\sum_{k=1}^{t} d_k^2 = \eta^2 \sum_{k=1}^{t} \frac{a_k}{A_{k-1}} \leq 2(\eta^2 + 4D^2)\log\frac{S_{t+1}}{\gamma}.$$

Applying Hölder's inequality to $\{d_k^{1+\nu}\}$ and $\{1\}$ with conjugate exponents $p = 2/(1+\nu)$ and $q = 2/(1-\nu)$, respectively, we obtain from the above inequality that

$$\sum_{k=1}^{t} d_k^{1+\nu} \leq t^{\frac{1-\nu}{2}}\left(\sum_{k=1}^{t} d_k^2\right)^{\frac{1+\nu}{2}} \lesssim t^{\frac{1-\nu}{2}}\left(\log\frac{S_{t+1}}{\gamma}\right)^{\frac{1+\nu}{2}}. \tag{F.2}$$

To proceed, rearranging (B.2) gives

$$\langle g_k, x_{k+1} - x_k\rangle + h(x_{k+1}) - h(x_k) \leq -\frac{S_k}{\eta} d_k^2.$$

Therefore, applying (F.1) and using $\nabla f(x_k) = g_k + \delta_k$, we get

$$\begin{aligned}\Delta_{k+1} - \Delta_k &\leq \langle g_k, x_{k+1} - x_k\rangle + h(x_{k+1}) - h(x_k) + \langle \delta_k, x_{k+1} - x_k\rangle + \frac{L_\nu}{1+\nu} d_k^{1+\nu} \\ &\leq -\frac{S_k}{\eta} d_k^2 + \frac{L_\nu}{1+\nu} d_k^{1+\nu} + \langle \delta_k, x_{k+1} - x_k\rangle. \end{aligned} \tag{F.3}$$

Summing (F.3) over $k = 1, \ldots, t$ and rearranging then yields

$$2\eta(S_{t+1} - \gamma) \leq \sum_{k=1}^{t} \frac{S_k d_k^2}{\eta} \leq \Delta_1 + \frac{L_\nu}{1+\nu}\sum_{k=1}^{t} d_k^{1+\nu} + \sum_{k=1}^{t}\langle \delta_k, x_{k+1} - x_k\rangle,$$

where the first inequality uses $\sum_{k=1}^{t} S_k d_k^2/\eta = \eta\sum_{k=1}^{t}\|\widetilde{\mathcal{G}}_k\|^2/S_k \geq 2\eta(S_{t+1} - \gamma)$ from (A.1).

We next bound $m_t := \mathbb{E}[S_{t+1}/\gamma] \geq 1$. Taking expectations on both sides of the above inequality and applying (F.2) and Lemma 1 with $b_k = 1$ and $2D$ in place of D gives

$$\begin{aligned} m_t &\lesssim 1 + t^{\frac{1-\nu}{2}}\mathbb{E}\left[(\log(S_{t+1}/\gamma))^{\frac{1+\nu}{2}}\right] + \sigma\sqrt{t\log m_t} \\ &\lesssim 1 + t^{\frac{1-\nu}{2}}(\log m_t)^{\frac{1+\nu}{2}} + \sigma\sqrt{t\log m_t}, \end{aligned}$$

where the last inequality uses the concavity of $u \mapsto (\log u)^a$ for $a \in [0,1]$ on $[1,\infty)$. For simplicity, set

$$H_t := 1 + t^{\frac{1-\nu}{2}} + \sigma\sqrt{t}.$$

Then it follows from $\nu \in [0,1]$ that

$$m_t \leq CH_t(1 + \log m_t)^{\frac{1+\nu}{2}} \leq CH_t(1 + \log m_t) \tag{F.4}$$

for some constant $C > 0$. Applying $\log x \leq x - 1$ with $x = m_t/(2CH_t)$ to the above inequality gives

$$m_t \leq 2CH_t\log(2CH_t),$$

which then implies $\log m_t \lesssim 1 + \log H_t$ and hence we obtain from (F.4) that

$$m_t \lesssim H_t(1 + \log H_t)^{\frac{1+\nu}{2}}. \tag{F.5}$$

For the objective gap, convexity gives $f(x_k) - f(x^*) \le \langle \nabla f(x_k), x_k - x^* \rangle$, which together with (F.1) at $(x, y) = (x_k, x_{k+1})$ yields

$$f(x_{k+1}) - f(x^*) \le \langle \nabla f(x_k), x_{k+1} - x^* \rangle + \frac{L_\nu}{1+\nu} d_k^{1+\nu}.$$

Combining the above inequality with (C.5), $\nabla f(x_k) = g_k + \delta_k$, and $\langle \delta_k, x_{k+1} - x^* \rangle = \langle \delta_k, x_k - x^* \rangle + \langle \delta_k, x_{k+1} - x_k \rangle$ leads to

$$\begin{aligned}\Delta_{k+1} \le & \frac{S_k}{2\eta}\big(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - d_k^2\big) + \frac{L_\nu}{1+\nu} d_k^{1+\nu} \\ & + \langle \delta_k, x_k - x^* \rangle + \langle \delta_k, x_{k+1} - x_k \rangle.\end{aligned}$$

Summing the above inequality over $k = 1, \dots, t$ and taking expectations, we obtain from (C.15), $\mathbb{E}[\langle \delta_k, x_k - x^* \rangle] = 0$, (F.2), (F.5), and Lemma 1 with $b_k = 1$ and $2D$ in place of D that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^t \Delta_{k+1}\Big] \le & \frac{D^2 \mathbb{E}[S_{t+1}]}{2\eta} + \frac{L_\nu}{1+\nu} \sum_{k=1}^t \mathbb{E}[d_k^{1+\nu}] + \sum_{k=1}^t \mathbb{E}[\langle \delta_k, x_{k+1} - x_k \rangle] \\ \lesssim & m_t + t^{\frac{1-\nu}{2}} (\log m_t)^{\frac{1+\nu}{2}} + \sigma \sqrt{t \log m_t} \\ \lesssim & H_t (1 + \log H_t)^{\frac{1+\nu}{2}}.\end{aligned}$$

Finally, the result follows from the convexity of F and $H_t/t \lesssim t^{-(1+\nu)/2} + \sigma/\sqrt{t}$, which completes the proof. □

F.2 Convergence of Algorithm 2 under Hölder smoothness

Theorem 8. *Suppose f is convex and (L_ν, ν)-Hölder smooth for some $\nu \in [0, 1]$, and Assumptions 2 and 3 hold. Let $\{y_k\}_{k \ge 1}$ be generated by Algorithm 2 with $b_k \equiv 1$, $\gamma > 0$, and $\eta > \sqrt{2}D/2$. Then for any $t \ge 1$, we have*

$$\mathbb{E}[F(y_{t+1}) - F^*] \lesssim \Big(t^{-\frac{1+3\nu}{2}} + \frac{\sigma}{\sqrt{t}}\Big)\big[1 + \log\big(1 + t^{\frac{3(1-\nu)}{2}} + \sigma t^{3/2}\big)\big]^{\frac{1+\nu}{2}}.$$

Proof. For the cases $\nu = 0$ and $\nu = 1$, the result follows directly from Theorem 5 and Theorem 6, respectively. Hereafter, we assume $0 < \nu < 1$.

Recall that $\Delta_k := F(y_k) - F^*$ and $\delta_k := \nabla f(x_k) - g_k$. For ease of notation, let $d_k := \|z_{k+1} - z_k\|$ and $W_t := \sum_{k=1}^t \alpha_k^2$. Since $S_{k+1}^2 = S_k^2(1 + d_k^2/\eta^2)$, with $d_k \le 2D$ by Assumption 3, we have from (A.2) in Lemma 5 with $a_k = S_k^2 d_k^2/\eta^2$, $A_k = S_{k+1}^2$, and $C = 4D^2/\eta^2$ that

$$\sum_{k=1}^t d_k^2 = \eta^2 \sum_{k=1}^t \frac{a_k}{A_{k-1}} \le 2(\eta^2 + 4D^2) \log \frac{S_{t+1}}{\gamma}.$$

Applying Hölder's inequality to $\{\alpha_k^{1-\nu}\}$ and $\{d_k^{1+\nu}\}$ with conjugate exponents $p = 2/(1-\nu)$ and $q = 2/(1+\nu)$ respectively then gives

$$\sum_{k=1}^t \alpha_k^{1-\nu} d_k^{1+\nu} \le W_t^{\frac{1-\nu}{2}} \left(\sum_{k=1}^t d_k^2\right)^{\frac{1+\nu}{2}} \lesssim W_t^{\frac{1-\nu}{2}} \left(\log \frac{S_{t+1}}{\gamma}\right)^{\frac{1+\nu}{2}}. \tag{F.6}$$

Applying (F.1) at (x_k, y_{k+1}) and using $y_{k+1} - x_k = \theta_k(z_{k+1} - z_k)$ and (D.5) with $x = x^*$ leads to

$$\begin{aligned}f(y_{k+1}) & \le f(x_k) + \theta_k \langle \nabla f(x_k), z_{k+1} - z_k \rangle + \frac{L_\nu}{1+\nu} \theta_k^{1+\nu} d_k^{1+\nu} \\ & \le (1 - \theta_k) f(y_k) + \theta_k f(x^*) + \theta_k \langle \nabla f(x_k), z_{k+1} - x^* \rangle + \frac{L_\nu}{1+\nu} \theta_k^{1+\nu} d_k^{1+\nu}.\end{aligned}$$

Substituting $\nabla f(x_k) = g_k + \delta_k$ into the above inequality, applying (D.10) with $x = x^*$, and multiplying both sides by α_k^2, we obtain

$$\begin{aligned}\alpha_k^2\Delta_{k+1} \le& \alpha_{k-1}^2\Delta_k + \frac{S_k}{2\eta}\left(\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2 - d_k^2\right) + \frac{L_\nu}{1+\nu}\alpha_k^{1-\nu}d_k^{1+\nu} \\ &+ \alpha_k\langle\delta_k, z_{k+1} - z_k\rangle + \alpha_k\langle\delta_k, z_k - x^*\rangle,\end{aligned}$$

where we used $\theta_k = 1/\alpha_k$ and $\alpha_k^2 - \alpha_k = \alpha_{k-1}^2$ to simplify the result. Summing this inequality over $k = 1, \dots, t$, we further obtain from (D.15) and $\sum_{k=1}^t S_k d_k^2/(2\eta) = \eta\sum_{k=1}^t \|\widetilde{\mathcal{G}}_{z,k}\|^2/(2S_k) \ge \eta(S_{t+1} - \gamma)$ due to (A.1) that

$$\begin{aligned}\alpha_t^2\Delta_{t+1} + \left(\eta - \frac{D^2}{2\eta}\right)S_{t+1} \le& \eta\gamma + \frac{L_\nu}{1+\nu}\sum_{k=1}^t \alpha_k^{1-\nu}d_k^{1+\nu} \\ &+ \sum_{k=1}^t \alpha_k\langle\delta_k, z_{k+1} - z_k\rangle + \sum_{k=1}^t \alpha_k\langle\delta_k, z_k - x^*\rangle. \qquad \text{(F.7)}\end{aligned}$$

Let $m_t := \mathbb{E}[S_{t+1}/\gamma] \ge 1$. Note that the coefficient of S_{t+1} in (F.7) is positive due to $\eta > D/\sqrt{2}$ and that the last sum there has zero expectation. Therefore, taking expectations on both sides of (F.7) and applying (F.6) and Lemma 10 gives

$$m_t \lesssim 1 + W_t^{\frac{1-\nu}{2}}(\log m_t)^{\frac{1+\nu}{2}} + \sigma\sqrt{W_t \log m_t}.$$

For simplicity, set

$$H_t := 1 + W_t^{\frac{1-\nu}{2}} + \sigma\sqrt{W_t}.$$

Then it follows from the same argument as in the proof of Theorem 7 that

$$m_t \lesssim H_t(1 + \log H_t)^{\frac{1+\nu}{2}}.$$

To bound $\mathbb{E}[\Delta_{t+1}]$, we drop the nonnegative S_{t+1} term in (F.7) and take expectations to obtain

$$\alpha_t^2\mathbb{E}[\Delta_{t+1}] \lesssim 1 + W_t^{\frac{1-\nu}{2}}(\log m_t)^{\frac{1+\nu}{2}} + \sigma\sqrt{W_t \log m_t} \lesssim H_t(1 + \log H_t)^{\frac{1+\nu}{2}}. \qquad \text{(F.8)}$$

Finally, (D.12) gives $\alpha_t \ge (t+1)/2$ and $W_t \le t(t+1)^2 \le 4t^3$, which further implies

$$\frac{H_t}{\alpha_t^2} \lesssim t^{-\frac{1+3\nu}{2}} + \frac{\sigma}{\sqrt{t}}, \qquad \log H_t \lesssim 1 + \log(1 + t^{\frac{3(1-\nu)}{2}} + \sigma t^{3/2}).$$

Then the result follows from dividing both sides of (F.8) by α_t^2, which completes the proof. □

G Additional Experiments

This section provides additional experimental results that complement Section 6.

G.1 Regularized logistic regression

We evaluate the performance of Algorithms 1 and 2 on a convex composite problem: binary logistic regression with ℓ_1 regularization and box constraints,

$$\min_{x\in[-R,R]^d} F(x) := \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-b_i x^\top a_i\right)\right) + \lambda\|x\|_1. \qquad \text{(G.1)}$$

Here, $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$ denotes the i-th sample with feature vector $a_i \in \mathbb{R}^d$ and label $b_i \in \{-1, 1\}$. This problem fits into the composite formulation (1.1) by defining

$$f := \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-b_i\langle\cdot, a_i\rangle\right)\right), \quad h := \lambda\|\cdot\|_1 + \iota_{[-R,R]^d}(\cdot).$$

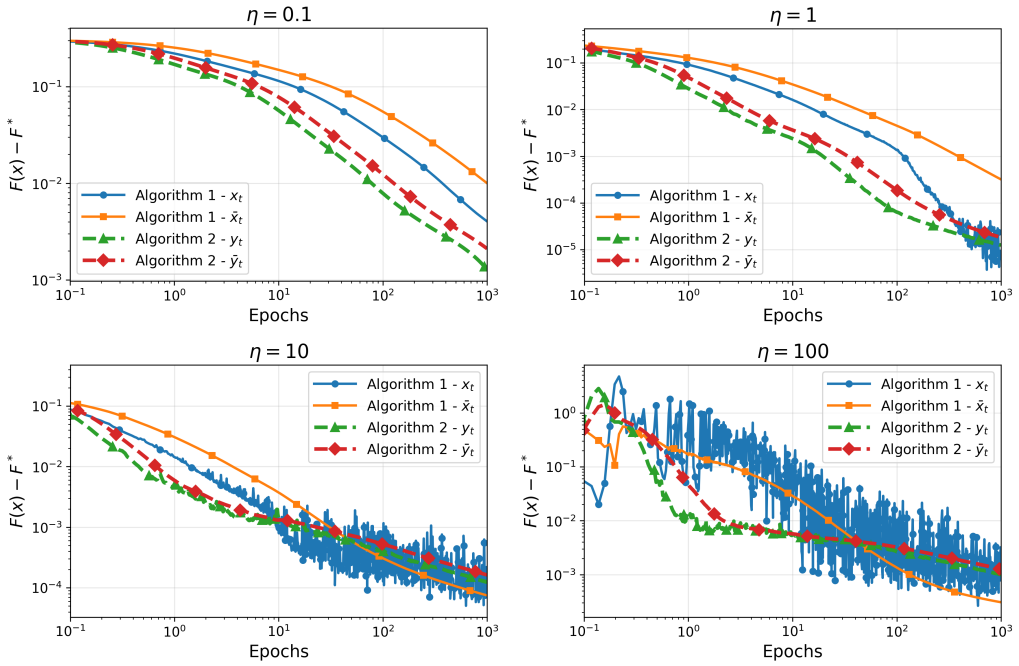


Figure 3: Comparison between Algorithms 1 and 2 for the regularized logistic regression problem (G.1) on the w6a dataset under different values of η. The averaged iterates are defined as $\bar{x}_t := \frac{1}{t}\sum_{k=1}^{t} x_{k+1}$ for Algorithm 1 and $\bar{y}_t := \frac{1}{\sum_{k=1}^{t}\alpha_k}\sum_{k=1}^{t}\alpha_k y_k$ for Algorithm 2.

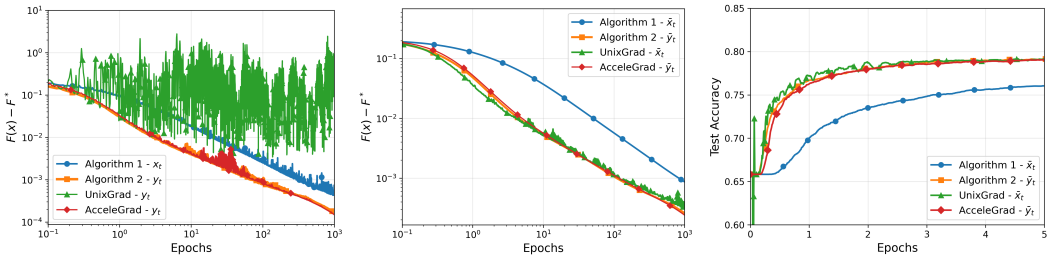
For the experiments on (G.1), we set $R = 50$. Since the objective function F is convex, we evaluate optimality using the suboptimality gap $F(x) - F^*$, where the optimal objective value F^* is approximated using the Python package CVXPY.

Comparison between Algorithms 1 and 2. We first set the ℓ_1 regularization parameter to $\lambda = 10^{-3}$ and compare the convergence behavior of Algorithms 1 and 2 on the w6a dataset [6] using a fixed batch size of 512. To assess robustness with respect to hyperparameter choices, we fix $\gamma = 1$ and consider four values of $\eta \in \{0.1, 1, 10, 100\}$ for both algorithms. Recall that Theorem 4 establishes convergence at the averaged iterate for Algorithm 1, whereas Theorem 6 guarantees last-iterate convergence for Algorithm 2. This motivates evaluating performance at both the last and averaged iterates.

The results are presented in Figure 3. We observe that Algorithm 1 achieves stable convergence in terms of the averaged iterate across all choices of η, but its last-iterate performance degrades as η increases. In contrast, Algorithm 2 exhibits robust convergence at both the last and averaged iterates, with similar convergence behavior across different values of η. These observations are consistent with Theorems 4 and 6.

Comparison in Non-Composite Settings. We further compare Algorithms 1 and 2 with existing adaptive accelerated methods UniXGrad [16] and AcceleGrad [26]. Since the latter two methods are designed for non-composite problems, we focus on the special case $\lambda = 0$. To conduct the experiments, we use the connect-4 dataset [6] and split it into training (80%) and test (20%) sets. For all methods, we set the batch size to $b = 512$ and run the algorithms for 1000 epochs. We fix $\eta = 10$ for Algorithms 1 and 2, and use the same initial step sizes for UniXGrad and AcceleGrad to ensure a fair comparison. All other settings are kept consistent with the previous experiment.

Figure 4 reports the convergence at both the last and averaged iterates, as well as test accuracies for the four methods. We observe that Algorithm 2 and AcceleGrad achieve comparable and strong performance in both convergence and test accuracy. In contrast, Algorithm 1 converges more slowly and yields lower test accuracy in the early stages. Moreover, UniXGrad exhibits fast convergence in terms of averaged iterates and improves test accuracy quickly, but its last-iterate behavior is highly unstable.



(a) Last iterate (b) Averaged iterate (c) Test accuracy

Figure 4: Comparison among Algorithm 1, Algorithm 2, UniXGrad, and AcceleGrad for solving the logistic regression problem (G.1) with $\lambda = 0$ on the connect-4 dataset.

Table 2: Selected hyperparameters for the full-gradient comparisons. PG step sizes are in units of $1/L$.

		Nonconvex SVM (6.1)			Logistic regression (G.1)		
Method	Parameter	MNIST	a9a	w6a	w6a	a9a	connect-4
Algorithm 1	η	10	10	100	100	100	1000
Algorithm 2	η	10	10	100	100	10	10
AdaGrad-Norm	η	10	10	100	100	100	100
AdaGrad (diag.)	η	10	10	100	100	10	10
UniSgd	$\hat{D}$	10	10	100	100	100	100
UniFastSgd	$\hat{D}$	–	–	–	100	10	10
PG	αL	1	10	10	10	1	1

G.2 Full-gradient comparisons

Setup. We solve the nonconvex SVM problem (6.1) with $\mu = \lambda = 10^{-3}$ on MNIST, a9a, and w6a, and logistic regression (G.1) with $\lambda = 10^{-3}$ on w6a, a9a, and connect-4 [6], using $R = 50$, $\gamma = 1$, and $x_1 = 0$. For MNIST, we use the binary labels of Section 6, and for connect-4, we assign the label $+1$ to wins and -1 to losses and draws. All methods use full gradients for 5000 iterations. We report the relative gradient mapping norm $\|\mathcal{G}(x_t)\|/\|\mathcal{G}(x_1)\|$ for (6.1) and the relative objective gap $(F(x_t) - F^*)/(F(x_1) - F^*)$ for (G.1), at the last iterate (y_t for Algorithm 2). We estimate F^* from a solution satisfying $\|\mathcal{G}(\cdot)\| < 10^{-9}$.

Methods. We compare Algorithms 1 and 2 with AdaGrad-Norm and coordinate-wise AdaGrad [8] (denoted AdaGrad (diag.) in the tables and figures), UniSgd and UniFastSgd [38], DoG [14], DoWG [18], and PG, all with proximal steps. UniSgd and UniFastSgd use the same initial step-size grid as our methods, parameterized by $\hat{D}/\gamma$ with diameter estimate $\hat{D}$. UniFastSgd is used only for logistic regression. For DoG and DoWG, we set the initial distance estimate $r_\epsilon = 10^{-6}$.

Hyperparameters. We use $\eta \in \{0.1, 1, 10, 100, 1000\}$ for our methods and both AdaGrad variants, the same grid for $\hat{D}$, and $\alpha \in \{0.01, 0.1, 1, 10\}/L$ for PG. Here, $L = \frac{4}{3\sqrt{3}}\lambda_{\max}(A^\top A/n) + \mu$ for (6.1) and $L = \frac{1}{4}\lambda_{\max}(A^\top A/n)$ for (G.1), with A the feature matrix. For hyperparameter tuning, we select the value reaching relative accuracy 10^{-6} with the fewest gradient evaluations, or the smallest final relative error if none reaches it. Table 2 lists the selected values.

Results. Tables 3–4 and Figures 5–6 show that Algorithm 1 is competitive with the AdaGrad variants and UniSgd, and reaches 10^{-6} fastest on connect-4. DoWG and PG perform best on several other datasets. Acceleration helps at moderate accuracies in logistic regression but does not improve the highest-accuracy results, while Algorithm 2 stalls in the nonconvex runs (see also Remark 4).

G.3 Constant and increasing batch sizes

Setup. We run Algorithm 1 with $\eta = 10$ and $\gamma = 1$ for 200 epochs on the nonconvex SVM problem (6.1) with a9a and the logistic regression problem (G.1) with w6a, using the accuracy

Table 3: Full-gradient SVM (6.1): gradient evaluations to reach $\|\mathcal{G}(x_t)\|/\|\mathcal{G}(x_1)\| \leq \epsilon$ and time (seconds) to reach $\epsilon = 10^{-6}$, with the hyperparameters from Table 2. Bold marks the best entry in each column; "–" means the target is not reached within 5000 iterations.

	MNIST				a9a				w6a			
Method	10^{-2}	10^{-4}	10^{-6}	Time	10^{-2}	10^{-4}	10^{-6}	Time	10^{-2}	10^{-4}	10^{-6}	Time
Algorithm 1	115	495	934	2.91	**13**	406	850	0.28	**13**	32	62	**0.02**
Algorithm 2	425	–	–	–	49	–	–	–	34	1574	–	–
AdaGrad-Norm	118	518	1005	3.13	14	407	860	0.28	**13**	32	62	**0.02**
AdaGrad (diag.)	115	497	940	2.93	17	401	851	0.28	**13**	32	**60**	**0.02**
UniSgd	115	488	920	2.87	14	406	849	0.28	14	34	63	**0.02**
DoG	366	1033	1827	5.69	703	2008	3860	1.27	264	1246	2540	0.92
DoWG	**103**	**270**	**500**	**1.56**	36	**150**	**249**	**0.08**	38	70	122	0.04
PG	355	1530	2889	8.99	22	161	300	0.10	**13**	**31**	**60**	**0.02**

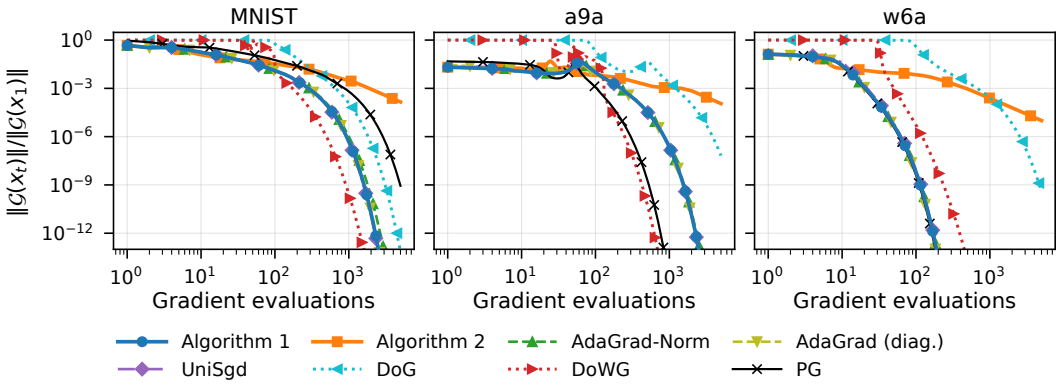


Figure 5: Nonconvex SVM (6.1) with full gradients: relative gradient mapping norm versus gradient evaluations, with the hyperparameters of Table 2.

Table 4: Full-gradient logistic regression (G.1): gradient evaluations to reach $(F(x_t)-F^*)/(F(x_1)-F^*) \leq \epsilon$ and time (seconds) to reach $\epsilon = 10^{-6}$, with the hyperparameters from Table 2. Bold marks the best entry in each column; "–" means the target is not reached within 5000 iterations.

	w6a				a9a				connect-4			
Method	10^{-2}	10^{-4}	10^{-6}	Time	10^{-2}	10^{-4}	10^{-6}	Time	10^{-2}	10^{-4}	10^{-6}	Time
Algorithm 1	10	75	173	0.07	879	896	1041	0.39	611	830	**1262**	**1.18**
Algorithm 2	**9**	79	812	0.31	**44**	397	3988	1.49	**66**	**331**	2726	2.40
AdaGrad-Norm	10	75	174	0.06	831	848	998	0.38	2045	2399	3675	3.42
AdaGrad (diag.)	10	75	173	0.06	216	1122	2257	0.82	528	3592	–	–
UniSgd	11	76	174	0.07	237	**264**	530	0.20	539	1260	–	–
UniFastSgd	18	144	1440	0.53	90	784	7828	2.82	118	502	4298	3.72
DoG	284	1139	2390	0.90	558	1850	3391	1.26	988	3523	–	–
DoWG	48	79	151	0.06	96	277	**491**	**0.18**	257	828	2311	2.07
PG	13	**26**	**55**	**0.02**	229	1197	2413	0.88	951	–	–	–

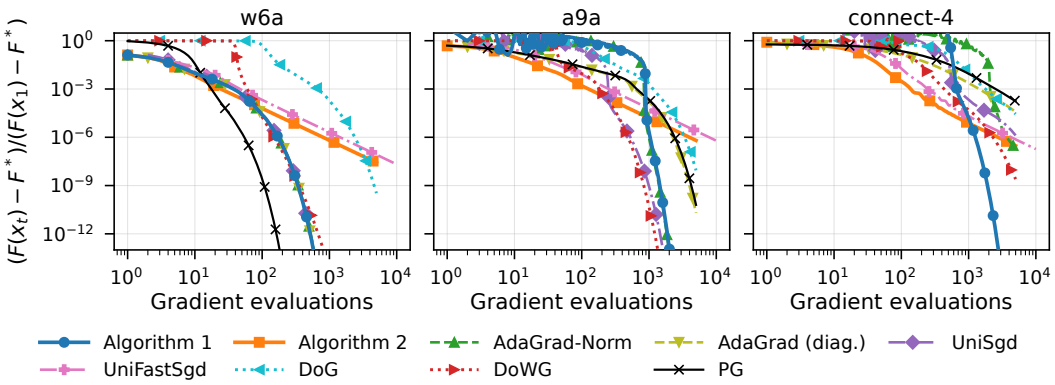


Figure 6: Logistic regression (G.1) with full gradients: relative objective gap versus gradient evaluations, with the hyperparameters of Table 2.

measures of Appendix G.2 for both problems. One epoch corresponds to n per-sample gradient evaluations. We use the constant batch sizes $b \in \{0.1\%, 1\%, 10\%, 100\%\}$ of n and the increasing batch sizes $b_k = \min\{n, \lceil b_1 \rho^{k-1} \rceil\}$ with $b_1 = 0.1\%$ of n and $\rho \in \{1.01, 1.05\}$. We compare with UniSgd and DoWG at $b = 0.1\%$ of n. UniSgd's $\hat{D}$ is selected on the first seed using the grid and rule of Appendix G.2, giving 0.1 for SVM and 1 for logistic regression. All methods reshuffle the data each epoch.

Results. Figure 7 shows that constant batch sizes $b < n$ leave a residual error for the nonconvex SVM, while increasing batch sizes reduce it substantially. For logistic regression, constant mini-batches continue to improve the objective, with $b = 10\%$ giving the smallest final relative gap.

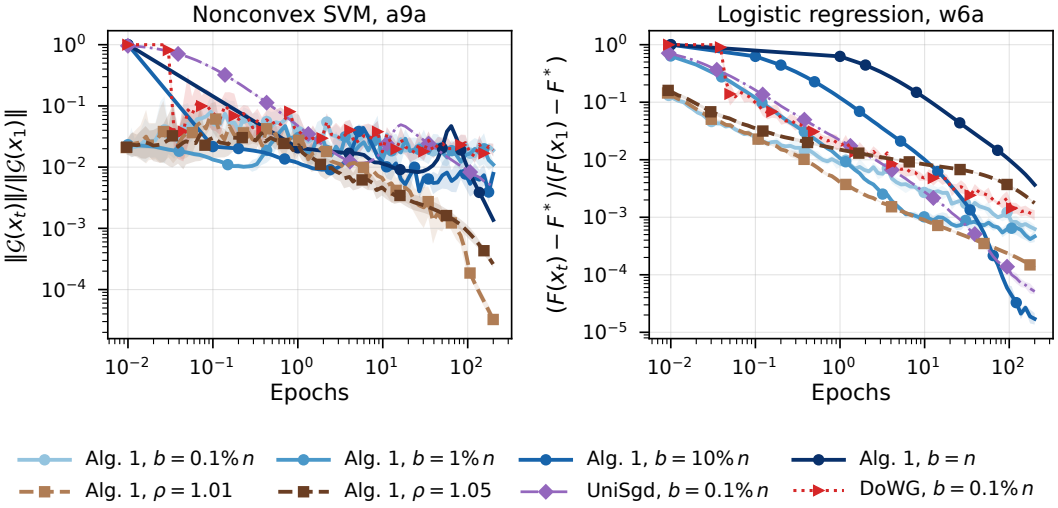


Figure 7: Constant and increasing batch sizes for SVM on a9a (left) and logistic regression on w6a (right), with UniSgd and DoWG as baselines. Curves show geometric means over 5 seeds, with shaded bands showing standard deviation.

G.4 Sensitivity to η and γ

We test Algorithm 1, Algorithm 2, and AdaGrad-Norm on w6a for 2000 full-gradient iterations, using $\eta \in \{10^{-2}, 10^{-1.5}, \ldots, 10^{2}\}$ and $\gamma \in \{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$. Figure 8 shows similar sensitivity for Algorithm 1 and AdaGrad-Norm: small η/γ slows progress, while $\gamma = 1$ and $\eta \geq 10$ perform well in the tested range. Algorithm 2 makes limited progress on SVM but is less sensitive to η on logistic regression.

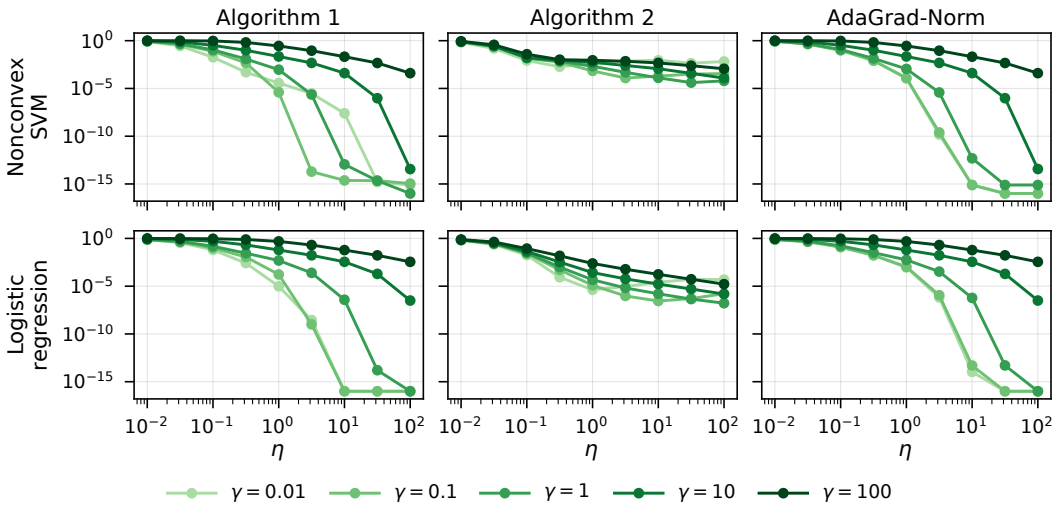


Figure 8: Sensitivity to η and γ on w6a after 2000 full-gradient iterations: relative gradient mapping norm for SVM (top) and relative objective gap for logistic regression (bottom).

G.5 Algorithm 2 with $\eta \leq D/\sqrt{2}$

Recall that our analysis requires $\eta > D/\sqrt{2}$ in Theorems 5 and 6. To show that the algorithm can still work in practice without this condition, we run Algorithm 2 on w6a logistic regression with $\gamma = 1$ and $\eta \in \{10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$, using 5000 full-gradient iterations or 50 epochs with $b = 1\%$ of n. We use the box $[-3.5, 3.5]^d$, which contains the reference solution, and take its diameter $D = 7\sqrt{d} \approx 121.2$ as a valid bound in Assumption 3. The resulting sufficient threshold is $D/\sqrt{2} \approx 85.7$. Figure 9 shows convergence below this threshold, including at $\eta = 1$ and 10. Smaller values of η may slow down the convergence, while moderate values perform better with mini-batches. This is consistent with Remark 3.

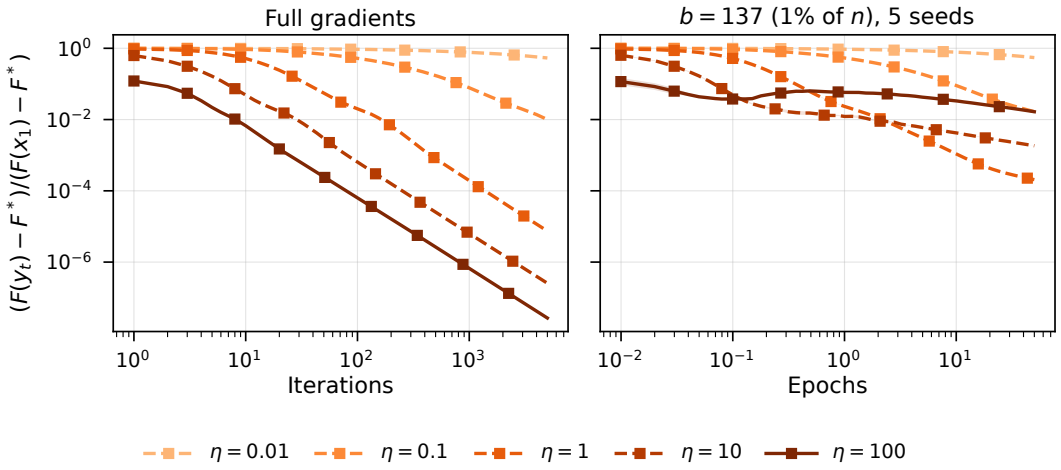


Figure 9: Algorithm 2 on w6a logistic regression with full gradients (left) and $b = 1\%$ of n (right). Dashed curves use $\eta \leq D/\sqrt{2}$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the convergence results for the main algorithm and the accelerated variant. All claims are formally established in Sections 3–5 and match the experimental evidence in Section 6.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 explicitly discusses several limitations of the work, including the logarithmic factors in the convergence rates, the boundedness assumptions, and the hyperparameter condition for the accelerated method.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All assumptions are rigorously stated in separate blocks and specified clearly in each theorem and corollary. Complete proofs for all theoretical results are provided in the appendix.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 6 and Appendix G fully specify the datasets, preprocessing steps, problem formulations, hyperparameter values, batch sizes, etc. used in all experiments.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code is provided in the supplemental material. All datasets used are publicly available and properly cited.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 6 and Appendix G report all hyperparameter choices, data splits, batch sizes, number of epochs, and performance metrics for each experiment.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Appendix G.3 and Figure 7 report geometric means and bands of one geometric standard deviation over five sampling seeds.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All experiments were run on an Apple M4 Pro CPU. The experiments are small-scale and can be reproduced on a standard CPU in minutes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This paper conducted experiments on publicly available datasets to test optimization algorithms. It poses no ethical concerns.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This is a theoretical paper on optimization algorithms. It has no direct path to negative societal impacts, and any positive impacts (such as improved training efficiency for ML models) are indirect and generic.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper uses only publicly available datasets and models, and therefore poses no such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The datasets used in the paper are publicly available and are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, models, or other releasable assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper involves neither crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology. They were used only for writing assistance, which does not require declaration per NeurIPS policy.

    Guidelines:

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