Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD

Matthew Faw* MATTHEWFAW@UTEXAS.EDU
Litu Rout* LITU.ROUT@UTEXAS.EDU
Constantine Caramanis CONSTANTINE@UTEXAS.EDU
Sanjay Shakkottai SANJAY.SHAKKOTTAI@UTEXAS.EDU
The University of Texas at Austin

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

This work considers the problem of finding a first-order stationary point of a non-convex function with potentially unbounded smoothness constant using a stochastic gradient oracle. We focus on the class of (L_0, L_1)-smooth functions proposed by Zhang et al. (ICLR'20). Empirical evidence suggests that these functions more closely capture practical machine learning problems as compared to the pervasive L_0-smoothness. This class is rich enough to include highly non-smooth functions, such as $\exp(L_1 x)$ which is $(0, \mathcal{O}(L_1))$-smooth. Despite the richness, an emerging line of works achieves the $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate of convergence when the noise of the stochastic gradients is deterministically and uniformly bounded. This noise restriction is not required in the L_0-smooth setting, and in many practical settings is either not satisfied, or results in weaker convergence rates with respect to the noise scaling of the convergence rate.

We develop a technique that allows us to prove $\mathcal{O}(\mathrm{poly}\log(T)/\sqrt{T})$ convergence rates for (L_0, L_1)-smooth functions without assuming uniform bounds on the noise support. The key innovation behind our results is a carefully constructed stopping time τ which is simultaneously "large" on average, yet also allows us to treat the adaptive step sizes before τ as (roughly) independent of the gradients. For general (L_0, L_1)-smooth functions, our analysis requires the mild restriction that the multiplicative noise parameter $\sigma_1 < 1$. For a broad subclass of (L_0, L_1)-smooth functions, our convergence rate continues to hold when $\sigma_1 \geq 1$. By contrast, we prove that many algorithms analyzed by prior works on (L_0, L_1)-smooth optimization diverge with constant probability even for smooth and strongly-convex functions when $\sigma_1 > 1$.

1. Introduction

A fundamental problem in stochastic optimization is to characterize the convergence behavior of the Stochastic Gradient Descent algorithm:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{g}(\mathbf{w}_t), \tag{SGD}$$

where η_t is the step-size schedule, and $\boldsymbol{g}(\mathbf{w}_t)$ is a stochastic gradient at iterate $\mathbf{w}_t$. Starting from (Robbins and Monro, 1951), a long line of work has established conditions under which (SGD) converges to a stationary point. A standard setting since (Polyak and Tsypkin, 1973) used for this purpose has the following properties: (a) The objective function $F(\cdot)$ is L_0-smooth, i.e., has L_0-Lipschitz gradients; (b) $F(\cdot)$ has a finite lower bound, i.e., $\inf_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) \geq F^* > -\infty$; (c) For

* Equal contribution

each $\mathbf{w} \in \mathbb{R}^d$, the stochastic gradient $\boldsymbol{g}(\mathbf{w})$ is unbiased and has variance scaling at most affinely with $\|\nabla F(\mathbf{w})\|^2$, i.e.,

$$\mathbb{E}\left[\boldsymbol{g}(\mathbf{w})\right] = \nabla F(\mathbf{w}) \quad \text{and} \quad \mathbb{E}\left[\|\boldsymbol{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|^2\right] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(\mathbf{w})\|^2 . \tag{Affine-var}$$

Much of the literature on stochastic optimization, e.g., (Nemirovski and Yudin, 1983; Ghadimi and Lan, 2013; Bubeck, 2015; Foster et al., 2019), focuses on a special case of (Affine-var) where the variance is uniformly upper-bounded ($\sigma_1 = 0$):

$$\mathbb{E}\left[\boldsymbol{g}(\mathbf{w})\right] = \nabla F(\mathbf{w}) \quad \text{and} \quad \sup_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}\left[\|\boldsymbol{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|^2\right] \leq \sigma_0^2. \tag{Bounded-var}$$

Rates of convergence to a first-order stationary point in these settings are now well-understood. Under (Bounded-var) regime, Ghadimi and Lan (2013) prove an $\mathcal{O}(\sqrt{\sigma_0^2 L_0 (F(\mathbf{w}_1) - F^*)}/\sqrt{T})$ rate of convergence with a fixed step-size schedule. Later, Arjevani et al. (2022) show that this rate is optimal up to constant factors. Further, as noted by Bottou et al. (2018), a minor modification to this step-size gives nearly the same rate in the more general (Affine-var) setting, i.e., $\sigma_1 > 0$. This rate is obtained by making trivial changes to the proof technique of Ghadimi and Lan (2013).

One crucial assumption in these lines of work is L_0-smoothness, i.e., L_0-Lipschitz gradients of the loss landscape. However, recent works (Zhang et al., 2020a,b) provide empirical evidence that this assumption is often not satisfied in practical machine learning problems. For instance, in large-scale language modeling including BERT (Devlin et al., 2018) and other variants (Radford et al., 2021; Caron et al., 2021; Liu et al., 2023), the loss landscape of transformer architectures either does not satisfy the L_0-smoothness assumption, or the value of L_0 becomes so large that it produces a significantly weaker rate of convergence (Zhang et al., 2020a,b).

Aiming to address these issues, there has been a recent surge of interest in relaxing the standard L_0-smoothness assumption and characterizing the rate of convergence. One appealing relaxation proposed by Zhang et al. (2020b) is that of (L_0, L_1)-smoothness[1]:

$$\left\|\nabla^2 F(\mathbf{w})\right\| \leq L_0 + L_1 \|\nabla F(\mathbf{w})\| . \tag{Generalized-smooth}$$

While every L_0-smooth function is also $(L_0, 0)$-smooth, this relaxation admits functions that grow significantly faster than a quadratic function, e.g., $F(w) = w^d$ is $(d(d-1)/L_1^{d-2}, (d-1)L_1)$-smooth for any $L_1 > 0$, and $F(w) = \exp(L_1 w)$ is $(0, L_1)$-smooth. With regards to convergence, recent works (Zhang et al., 2020b; Crawshaw et al., 2022) establish an $\mathcal{O}(1/\sqrt{T})$ rate in the (L_0, L_1)-smooth setting, as long as the noise of the stochastic gradients has *uniformly-bounded support*, i.e.,

$$\mathbb{E}\left[\boldsymbol{g}(\mathbf{w})\right] = \nabla F(\mathbf{w}) \quad \text{and} \quad \sup_{\mathbf{w}\in\mathbb{R}^d} \|\boldsymbol{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \overset{a.s.}{\leq} B^2. \tag{Bounded-supp}$$

The algorithms achieving the rate $\mathcal{O}(1/\sqrt{T})$ in this setting use adaptive step size schedules – i.e., variants of (SGD) with η_t chosen as a function of $\{\boldsymbol{g}(\mathbf{w}_s)\}_{s\in[t]}$. (Bounded-supp) is a common assumption in these analyses (Zhang et al., 2020b,a; Crawshaw et al., 2022). It is typically introduced to reason about the direction of $-\eta_t \boldsymbol{g}(\mathbf{w}_t)$ relative to the true descent direction. In the analysis

1. For convenience, we state this assumption in terms of a bound on the hessian of F. The requirement that the hessian exists everywhere can be relaxed to a condition on the gradients (Zhang et al., 2020a). This relaxation is the one we use for our main results, see Assumption 2.

of standard SGD, the fixed step-size schedule does not depend upon the stochastic gradients, so $\mathbb{E}\left[-\eta_t \boldsymbol{g}(\mathbf{w}_t)\right] = -\eta_t \mathbb{E}\left[\nabla F(\mathbf{w}_t)\right]$. This, however, is not the case for adaptive methods, since η_t depends on $\boldsymbol{g}(\mathbf{w}_t)$. Thus, it is understandable why prior works (Zhang et al., 2020b,a) assume (Bounded-supp) to simplify this issue. Further, (Bounded-supp) is natural in settings where the stochastic gradients satisfy $\boldsymbol{g}(\mathbf{w}) = \nabla F(\mathbf{w}) + \xi$, where the random vector ξ has bounded support (or bounded second moment in the related setting of (Bounded-var)).

In many real-world scenarios, the (Bounded-supp) assumption does not hold. For instance, when running SGD in standard least-squares regression settings, the stochastic gradients have multiplicative noise, as noted in (Dieuleveut et al., 2017; Flammarion and Bach, 2017; Jain et al., 2018; Jofré and Thompson, 2019). Similar noise assumptions have also been considered, e.g., in convergence of stochastic proximal gradient methods (Rosasco et al., 2020), Hilbert-valued stochastic subgradient methods (Barty et al., 2007), and adaptive gradient methods (Faw et al., 2022). Moreover, multiplicative noise naturally arises in machine learning problems with (additive or multiplicative) feature noise (Loh and Wainwright, 2011; Hwang, 1986; Carroll et al., 2006). Thus, we believe that characterizing (L_0, L_1)-smooth functions under (Affine-var) is an important step in extending the theory of non-convex stochastic optimization beyond the standard L_0-smooth setting.

1.1. Contributions

A major challenge in the analysis of adaptive stochastic gradient descent is the correlation between the stochastic gradients and the step-size. Here, we develop a technique to simplify this challenge. Our key innovation is a *recursively-defined stopping time* which satisfies two crucial properties: (i) before the stopping time is reached, the step sizes behave roughly independently of the gradients, and (ii) on average, the stopping time is at least a constant fraction of the time horizon. As a consequence, instead of analyzing over the entire time horizon, we conduct the analysis over this sub-interval over which we exploit this convenient almost-independent property. This tool allows us to prove the first $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate of convergence for (L_0, L_1)-smooth functions beyond the (Bounded-supp) setting. Our main contributions are three-fold:

(a) **Convergence for (L_0, L_1)-smoothness when $\sigma_1 < 1$.** We show in Section 4 that AdaGrad-Norm converges at a rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$ when the stochastic gradient oracle satisfies (Affine-var) with $\sigma_0 \geq 0$ and $\sigma_1 \in [0, 1)$. This is the first convergence rate for any algorithm even under (Bounded-var) (i.e., $\sigma_1 = 0$) for general (L_0, L_1)-smooth optimization. Note that the scaling of this bound with T matches (up to poly-logarithmic factors) the best-known rate for L_0-smooth functions – with a minor caveat that $\sigma_1 < 1$ is not needed in the L_0-smooth setting. Also, we show that the rate improves to $\widetilde{\mathcal{O}}(1/T)$ in the "small variance" regime when $\sigma_0, \sigma_1 \to 0$ even without tuning the step-size.

(b) **Convergence for all σ_1.** We establish a sufficient condition under which AdaGrad-Norm converges at a rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$ when $\sigma_1 \geq 1$, see Section 5. This condition allows us to analyze a broad subset of (L_0, L_1)-smooth functions that includes all L_0-smooth functions as well as fixed-degree polynomials without any restrictions on σ_1. This simultaneously generalizes the result and simplifies a key proof technique of Faw et al. (2022) for L_0-smooth functions.

(c) **Negative results for known algorithms.** We prove a set of negative results in Section 6 for most algorithms analyzed under (L_0, L_1)-smoothness and (Bounded-supp). We construct an oracle for Clipped and Normalized SGD (Zhang et al., 2020b,a) and Sign SGD with Momentum (Crawshaw et al., 2022) that leads to failure with constant probability in a wide parameter regime. We also prove that AdaGrad-Norm can diverge with constant probability if the step-size is not carefully tuned in the

"large variance" regime for (L_0, L_1)-smooth functions. By contrast, no parameter tuning is needed in the L_0-smooth setting in this noise regime.

2. Related Works

Stochastic gradient descent. (SGD) has been well-studied for many decades (Robbins and Monro, 1951). Polyak and Tsypkin (1973) proved almost-sure convergence to a first-order stationary point of (SGD) for non-convex and L_0-smooth functions with $F(\mathbf{w}) \geq F^*$ with stochastic gradient oracle satisfying (a slightly weaker condition than) (Affine-var). Bertsekas and Tsitsiklis (2000) extended the result to a setting where $F(\mathbf{w})$ does not have a uniform lower-bound. Ghadimi and Lan (2013) proved that (SGD) with step-size $\eta_t = \eta = \min\left\{1/L_0, \sqrt{2(F(\mathbf{w}_1)-F^*)/(L_0\sigma_0^2 T)}\right\}$ achieves a convergence rate to a first-order stationary point of $\mathcal{O}\left(\sqrt{L_0\sigma_0^2(F(\mathbf{w}_1)-F^*)/T}\right)$, assuming L_0-smoothness and (Bounded-var). Drori and Shamir (2020) proved that this is the optimal rate for (SGD) without further assumptions. Recently, Arjevani et al. (2022) proved that the convergence rate of (Ghadimi and Lan, 2013) is optimal among *all first-order* methods, not just SGD.

AdaGrad step-sizes. This paper builds on a long line of work studying (variants of) the AdaGrad step size schedule introduced by Duchi et al. (2011); McMahan and Streeter (2010). In particular, we focus on the so-called AdaGrad-Norm step-size, which was introduced in Streeter and McMahan (2010). While these works focused on the setting of online convex optimization, Ward et al. (2020) demonstrated that AdaGrad-Norm converges at a rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$ in the context of L_0-smoothness, (Bounded-var), and M-Lipschitzness, i.e., $\sup_{\mathbf{w}\in\mathbb{R}^d} \|\nabla F(\mathbf{w})\| \leq M$. Around the same time, Li and Orabona (2019) proved that AdaGrad-Norm achieves an $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate without M-Lipschitzness. But their analysis needs tuning of the step-size with respect to the smoothness constant L_0. Later, Kavis et al. (2022) proved that AdaGrad-Norm converges at rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$ without tuning the step-size (as in Li and Orabona (2019)) or assuming M-Lipschitz objective (as in Ward et al. (2020)). However, their analysis holds only when the noise of the stochastic gradients is uniformly sub-Gaussian. In a concurrent work, Faw et al. (2022) proved that AdaGrad-Norm achieves $\widetilde{\mathcal{O}}(1/\sqrt{T})$ in a setting identical to standard SGD (i.e., L_0-smooth objective with stochastic gradients satisfying (Affine-var)), and without tuning the step-size with respect to L_0, σ_0, or σ_1. This work thus established that AdaGrad-Norm is parameter-free and enjoys nearly the same convergence rate as SGD in the standard non-convex setting.

(L_0, L_1)-smoothness in the (Bounded-supp) regime. Recent work by Zhang et al. (2020b) argued that the L_0-smoothness assumption is not realistic for many practical machine learning tasks, e.g., large-scale natural language processing using transformer architectures. Instead, they demonstrated that (L_0, L_1)-smooth functions (Generalized-smooth) better capture the loss landscape, and proved that the gradient clipping algorithm converges at a rate $\mathcal{O}(1/\sqrt{T})$ in the (Bounded-supp) regime. Zhang et al. (2020a) later proved convergences for a generalized class of gradient clipping algorithms. They used a slightly weaker definition of (L_0, L_1)-smoothness, which we use in Assumption 2. Very recently, Crawshaw et al. (2022) considered a "coordinate-wise" generalization of (L_0, L_1)-smoothness, and proved that a "generalized SignSGD" algorithm converges at a rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$. By contrast, they proved that gradient descent with fixed step-sizes must scale linearly in ML_1, where $M = \sup\{\|\nabla F(\mathbf{w})\| : F(\mathbf{w}) \leq F(\mathbf{w}_1)\}$ is the largest gradient in the sublevel set $F(\mathbf{w}) \leq F(\mathbf{w}_1)$. Interestingly, this line of work establishes that adaptive step-size schedules can avoid this dependence on M.

Stopping time arguments in optimization. Within the stochastic approximation literature, there have been a significant number of works using stopping times either as an analytical tool (Bertsekas and Tsitsiklis, 2000; Patel, 2022; Patel et al., 2022; Patel and Berahas, 2022), or as a part of the algorithm design (Sielken Jr, 1973; Stroup and Braun, 1982; Curtis and Scheinberg, 2020; Patel, 2022). Throughout the majority of these works, the stopping times are designed to test for closeness to a stationary point or for a sufficient decrease in objective. The main exceptions are (Patel et al., 2022; Patel and Berahas, 2022), where stopping times are instead used to determine when a local descent inequality can be applied. The stopping time used in our analysis (see Definition 9) serves a significantly different role – its main purpose is to effectively decorrelate the AdaGrad-Norm step-sizes from the gradients.

Concurrent work. In a concurrent work also appearing in COLT'23, Wang et al. (2023) establish, using different techniques, a $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate of convergence for AdaGrad-Norm under (L_0, L_1)-smoothness and (Affine-var) without the constraint of $\sigma_1 < 1$ or the alternative restriction in Definition 4. Their proof relies on a very interesting observation: the bias between the stochastic gradient and step-size can essentially be upper-bounded by an auxiliary function that allows for a telescoping cancellation. This leads to a descent lemma (a stronger version of Lemma 8) that holds over all times $t \in [T]$. Our main results (Theorems 3 and 5), by contrast, rely on a stopping-time argument which effectively allows us to decorrelate the step-sizes from the gradients (see (3)). This technique can enable one to obtain a convergence rate in other settings (such as Lemma 8) where the descent inequality might not always hold; instead, it could hold only over a (large enough) random subset of $[T]$.

3. Problem Setting

We are interested in finding a first-order stationary point of a non-convex function, given access to a stochastic gradient oracle, using (SGD). For compactness, let $\boldsymbol{g}_t := \boldsymbol{g}(\mathbf{w}_t)$. Our objective function $F(\mathbf{w})$ satisfies the following:

Assumption 1 (Lower-boundedness) *There exists an $F^* > -\infty$ such that $\inf_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) \geq F^*$.*

Assumption 2 ((L_0, L_1)-smooth objective) *The objective function $F(\mathbf{w})$ is (L_0, L_1)-smooth, i.e., for every $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ such that $\|\mathbf{w} - \mathbf{w}'\| \leq 1/L_1$*

$$\left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\right\| \leq \left(L_0 + L_1 \left\|\nabla F(\mathbf{w}')\right\|\right) \left\|\mathbf{w} - \mathbf{w}'\right\|.$$

We note that (L_0, L_1)-smoothness was originally defined in (Zhang et al., 2020b) as a bound on the Hessian of $F(\cdot)$, as in (Generalized-smooth). Following (Zhang et al., 2020a, Remark 2.3), we choose to adopt the alternative condition in Assumption 2 for two reasons. First, Assumption 2 is strictly weaker than L_0-smoothness, since $(L_0, 0)$-smoothness implies the gradients are L_0-Lipschitz. Second, whenever the objective is twice-differentiable, Assumption 2 implies (Generalized-smooth) (up to constant factors in the definitions of L_0 and L_1):

Proposition 1 *A function satisfying (L_0, L_1)-smoothness according to (Generalized-smooth) is also $(2L_0, (e-1)L_1)$-smooth according to Assumption 2. If $F(\cdot)$ is twice continuously differentiable and (L_0, L_1)-smooth according to Assumption 2, then it is also (L_0, L_1)-smooth according to (Generalized-smooth).*

Let $\mathcal{F}_t$ be the sigma-algebra generated by the interaction between the algorithm and stochastic gradient oracle for t rounds, i.e., $\mathcal{F}_t := \sigma\{\mathbf{w}_1, \boldsymbol{g}_1, \ldots, \mathbf{w}_t, \boldsymbol{g}_t, \mathbf{w}_{t+1}\}$. We impose the following conditions on the stochastic gradients:

Assumption 3 (Unbiased gradients) *The stochastic gradients satisfy* $\mathbb{E}[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}] = \nabla F(\mathbf{w}_t)$.

Assumption 4 (Affine variance) *There exist constants* $\sigma_0, \sigma_1 \geq 0$ *such that the variance of each stochastic gradient* $\boldsymbol{g}_t$ *is bounded above as:* $\mathbb{E}\left[\|\boldsymbol{g}_t - \nabla F(\mathbf{w}_t)\|^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(\mathbf{w}_t)\|^2$.

Assumptions 3 and 4 imply the following bound on the stochastic gradients in terms of the true gradient:

$$\mathbb{E}\left[\|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma_0^2 + (1+\sigma_1^2)\|\nabla F(\mathbf{w}_t)\|^2. \tag{1}$$

We are interested in studying algorithms which require as little hyper-parameter tuning as possible and, simultaneously, can handle potentially unbounded smoothness constant. To achieve this, we analyze AdaGrad-Norm (Streeter and McMahan, 2010), a step-size sequence η_t for (SGD) which, at each time t, depends on the current and past stochastic gradients $\{\boldsymbol{g}_s\}_{s\in[t]}$:

$$\eta_t = \frac{\eta}{b_t}, \quad \text{where} \quad b_t^2 = b_0^2 + \sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2 = b_{t-1}^2 + \|\boldsymbol{g}_t\|^2. \tag{AG-Norm}$$

As is increasingly common in the analysis of (variants of) (SGD) with adaptive step-sizes (Ward et al., 2020; Faw et al., 2022; Défossez et al., 2022), our analysis will rely on a "decorrelated" step-size $\tilde{\eta}_t$. The key property of $\tilde{\eta}_t$ is that it is independent of $\boldsymbol{g}_t$ when conditioned on the filtration $\mathcal{F}_{t-1}$.

Definition 2 (Decorrelated step-sizes) *For each step-size* η_t *at time* $t \geq 1$*, the decorrelated step size* $\tilde{\eta}_t$ *is defined to be* $\tilde{\eta}_t := \eta/\widetilde{b}_t$*, where* $\widetilde{b}_t^2 := b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2$*,* $b_0^2 > 0$*, and* $\|\widetilde{\nabla}_t\|^2 := \sigma_0^2 + \|\nabla F(\mathbf{w}_t)\|^2$.

This "decorrelated" step-size serves as a proxy in our analysis for the true step-size η_t. The main reason for its introduction is that, although $\mathbb{E}[\eta_t \boldsymbol{g}_t] \neq \mathbb{E}[\eta_t \nabla F(\mathbf{w}_t)]$ (since η_t depends on $\boldsymbol{g}_t$), the proxy satisfies $\mathbb{E}[\tilde{\eta}_t \boldsymbol{g}_t \mid \mathcal{F}_{t-1}] = \tilde{\eta}_t \nabla F(\mathbf{w}_t)$.

4. Convergence of AdaGrad-Norm on (L_0, L_1)-smooth functions

Our main results, Theorems 3 and 5, both establish $\widetilde{\mathcal{O}}(1/\sqrt{T})$ convergence rates for (AG-Norm) in the (L_0, L_1)-smooth regime under (Affine-var). Theorem 3 holds for any (L_0, L_1)-smooth function under a mild restriction that $\sigma_1 < 1$. It is easy to extend this result for $\sigma_1 \geq 1$ by computing mini-batch gradients with a batch size $B \approx \sigma_1^2$, refer Fact 19 for a proof. Despite the restriction of Theorem 3 to $\sigma_1 < 1$, we emphasize that, prior to our work, no proof of convergence even for the (Bounded-var) setting (i.e., $\sigma_1 = 0$) was known for a general class of (L_0, L_1)-smooth functions. Besides, Theorem 5 holds for all σ_1 and a subclass of (L_0, L_1)-smooth functions, i.e., excluding functions like $\exp(L_1 w)$.

Theorem 3 (Informal statement of Theorem 26) *Fix any constants* $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$ *such that* $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$*. Consider* (AG-Norm) *with any parameters* $\eta \leq 2\varepsilon'/5L_1$ *and* $b_0^2 > 0$*, running on an objective function satisfying Assumption 2, and given access to a stochastic gradient*

oracle satisfying Assumptions 3 and 4. Assuming that $\sigma_1 \leq (1 - (\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon'''))$*, then for any* $T \geq 1$ *and* $\delta' \in (0,1)$*, with probability at least* $1 - \delta'$*, the iterates of* (AG-Norm) *satisfy*

$$\begin{aligned}\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 \lesssim{}& \frac{\sigma_0}{(\delta')^2\sqrt{T}}h(T) + \frac{\sigma_1\sqrt{\sigma_0}}{(\delta')^{2.25}\sqrt{T}}h(T)^{3/2} + \frac{\sigma_1\sqrt{(1+\sigma_1^2)}}{(\delta')^{2.5}\sqrt{T}}h(T)^2 \\ &+ \frac{1}{(\delta')^2 T}h(T)^2 + \frac{b_0}{(\delta')^2 T}h(T) + \frac{\sigma_1\sqrt{b_0+\eta L_0}h(T)^{1.5}}{(\delta')^2 T}, \\ \textit{where}\quad h(T) \propto{}& \frac{1}{\varepsilon'''}\left(\frac{F(\mathbf{w}_1) - F^*}{\eta} + \frac{\varepsilon''\sigma_0}{1+\sigma_1^2} + \left(\frac{\sigma_0}{\varepsilon} + \eta L_0\right)\log(g(T))\right), \\ g(T) \propto{}& \frac{T(1+\sigma_1^2)\left(\frac{\sigma_0}{\varepsilon} + \eta L_0\right)}{\varepsilon'' b_0}.\end{aligned}$$

To extend our convergence proofs beyond $\sigma_1 < 1$, we consider a subclass of (L_0, L_1)-smooth functions which satisfy the following additional assumption:

Definition 4 *A function* $F(\cdot)$ *is* k-polynomially bounded *for* $k \geq 2$ *if* $\forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$*, then there are constants* $c_k \geq 1$ *and* $c_k', L_0 > 0$ *such that:*

$$\|\nabla F(\mathbf{w})\| - c_k \left\|\nabla F(\mathbf{w}')\right\| \leq \max\left\{c_k' \left\|\mathbf{w} - \mathbf{w}'\right\|^{k-1}, L_0 \left\|\mathbf{w} - \mathbf{w}'\right\|\right\}.$$

Notice that, whereas Assumption 2 is a local constraint on the objective, Definition 4 enforces a global polynomial growth constraint – thus ruling out such (L_0, L_1)-smooth functions as exponentials, while capturing a significantly broader class of functions than L_0-smoothness. We refer the interested reader to Proposition 28 for some properties of this class of functions. Using this definition, we are able to prove the following:

Theorem 5 (Informal statement of Corollary 32) *Fix any constants* $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$ *such that* $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$*. Consider* (AG-Norm) *with any parameters* $\eta \leq {}^{2\varepsilon'}/_{L_1(4+\sigma_1^2)}$ *and* $b_0^2 > 0$*, running on an objective function satisfying Assumption 2 and Definition 4 for some constants* $k \geq 2, c_k \geq 1, c_k' > 0$*, and given access to a stochastic gradient oracle satisfying Assumptions 3 and 4 for any* $\sigma_0, \sigma_1 \geq 0$*. Then, for any* $T \geq 1$ *and* $\delta' \in (0,1)$*, with probability at least* $1 - \delta' - \widetilde{\mathcal{O}}(1/T)$*, the iterates of* (AG-Norm) *satisfy*

$$\begin{aligned}\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 \lesssim{}& \frac{\sigma_0}{(\delta')^2\sqrt{T}}\widetilde{h}(T) + \frac{\sigma_1\sqrt{\sigma_0(1+c_k^2)}}{(\delta')^{2.25}\sqrt{T}}\widetilde{h}(T)^{3/2} + \frac{\sigma_1(1+c_k^2)\sqrt{(1+\sigma_1^2)}}{(\delta')^{2.5}\sqrt{T}}\widetilde{h}(T)^2 \\ &+ \frac{\sigma_1\sqrt{1+c_k^2}\sqrt[4]{(1+\sigma_1^2)c_{\mathrm{B1}}}\widetilde{h}(T)^{1.5}}{(\delta')^{2.25}T^{3/4}} \\ &+ \frac{\left(b_0 + \sqrt{(1+\sigma_1^2)c_{\mathrm{B1}}}\right)\widetilde{h}(T)}{(\delta')^2 T} + \frac{\sigma_1\sqrt{(1+c_k^2)(b_0+\eta L_0)}\widetilde{h}(T)^{1.5}}{(\delta')^2 T} + \frac{(1+c_k^2)\widetilde{h}(T)^2}{(\delta')^2 T}\end{aligned}$$

where $\widetilde{h}(T) = h(T) + {}^{\mathrm{comp}(T)}/_{\varepsilon'''\eta}$, *where* $h(T)$ *is the function defined in Theorem 3, and*

$$\mathrm{comp}(T) \propto \eta\sigma_1 c_k \left(\|\nabla F(\mathbf{w}_1)\| \, \ell_1(T) + (c_k'\eta^{k-1} + L_0\eta) \left({}^{4c_k^3\sigma_1}/_{\varepsilon'''}\right)^{k-1} \ell_k(T) \right)$$

$$c_{\mathrm{B1}} \propto (c_k'\eta^{k-1} + \eta L_0)^2 (1 + {}^{c_k^3\sigma_1}/_{\varepsilon'''})^{2k-1} \ell_{2k-1}(T) + c_k^2 \, \|\nabla F(\mathbf{w}_1)\|^2 \, (1 + {}^{c_k^3\sigma_1}/_{\varepsilon'''}) \ell_1(T)$$

$$\ell_k(T) \propto \left(\frac{(k+1)\sigma_1^2 \log\left(e + 8e \frac{\sigma_0^2 T^2 + (1+\sigma_1^2)(T^2 c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 + (c_k'\eta^{k-1} + \eta L_0)^2 T^{2k-1})}{b_0^2 \delta'} \right)}{(1 - (\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon'''))^2} \right)^k$$

There are several notable takeaways from the above results.

Noise adaptivity. Both Theorems 3 and 5 provide "noise-adaptive" convergence rates, in a sense that as $\sigma_0, \sigma_1 \to 0$, the convergence rates automatically improve from $\widetilde{\mathcal{O}}(1/\sqrt{T})$ to $\widetilde{\mathcal{O}}(1/T)$ without any additional hyperparameter tuning. To the best of our knowledge, (AG-Norm) is the first algorithm for (L_0, L_1)-smooth optimization with this property.

Less hyperparameter tuning. These rates hold without tuning the algorithm's parameters with respect to σ_0 or L_0, unlike all prior algorithms for (L_0, L_1)-smoothness that we are aware of (Zhang et al., 2020b,a; Crawshaw et al., 2022)[2]. Unlike in the L_0-smooth setting, however, (AG-Norm) requires some hyperparameter tuning. In a concurrent work, (Wang et al., 2023, Theorem 9) establishes that (AG-Norm) can diverge if $\eta > {}^{9\sqrt{5}}/_{2L_1}$. Further, as we prove in Lemma 34, η must also depend on σ_1, at least in the "large variance" regime. Indeed, we show that, when $\eta \geq {}^{1}/_{L_1\sqrt{\sigma_1}}$ and $\sigma_1 = \mathrm{poly}\log(T)$, then (AG-Norm) can diverge with constant probability. By contrast, no tuning is necessary for this algorithm to converge for L_0-smooth functions at the $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate in this noise regime.

Generalization of prior work. We remark that Theorem 5 strictly generalizes the result of (Faw et al., 2022) beyond the uniform L_0-smooth setting, since every L_0-smooth function satisfies Definition 4 with $k = 2$, $c_k = 1$, and $c_k' = L_0$. Further, our stopped analysis simplifies their "recursive improvement" technique (Faw et al., 2022, Lemma 13) which they used to prove that $\mathbb{E}\left[\sum_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2\right] = \widetilde{\mathcal{O}}(T)$, a key step in obtaining both their and our convergence guarantees. Their proof of this lemma crucially relied on the fact that, under L_0-smoothness, $\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}(t^2)$ deterministically, and $\mathcal{O}(t\log(t/\delta))$ with probability at least $1-\delta$. Indeed, they highlight this in *Step 2* of their proof sketch of Lemma 13 (p. 14) and in the proof of their Lemma 33. Our analysis circumvents these complications by analyzing the convergence only until a stopping time $\tau_{T+1}(\delta)$. As we show in Lemma 10 and (3), this time essentially allows us to decorrelate the step size from the gradients. As a result, we obtain essentially the same bound on the expected sum of gradients while completely sidestepping the "recursive improvement" argument or bounds on $\|\nabla F(\mathbf{w}_t)\|^2$ implied by L_0-smoothness.

5. Key technical ideas

As discussed earlier, the main technical tool we use to obtain our convergence rates in Theorems 3 and 5 is a recursively-defined stopping time. Before we are ready to define this time and discuss its utility, we first give a brief overview of the main initial steps of our analysis.

2. This feature, however, manifests into a worse dependence on L_1 unlike (Zhang et al., 2020a; Crawshaw et al., 2022). Interestingly, while their algorithms (in the (Bounded-supp) regime) do not need to be explicitly tuned with respect to L_1, they do require tuning with respect to T, L_0, and σ_0. Further, their results hold only for sufficiently large T (as determined, in part, by L_1).

The standard first step in the analysis of SGD-like algorithms for L_0-smooth non-convex optimization is to prove that, at least on average, each update makes sufficient progress. This argument typically relies on the following inequality for L_0-smooth functions: for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$,

$$F(\mathbf{w}') - F(\mathbf{w}) \leq \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{L_0}{2} \|\mathbf{w}' - \mathbf{w}\|^2.$$

This inequality is no longer true for (L_0, L_1)-smooth functions. Indeed, it is clearly not satisfied for all $\mathbf{w}, \mathbf{w}'$ on the $(0, L_1)$-smooth function $\exp(L_1 w)$. However, Zhang et al. (2020b,a) note that a similar variant holds "locally" for $\|\mathbf{w} - \mathbf{w}'\| \leq 1/L_1$ (see Lemma 17). Using this variant, we obtain the following inequality, which is our first tool for studying the convergence of (AG-Norm).

Lemma 6 *Fix any $\varepsilon, \varepsilon' \in (0,1)$. Suppose that $\eta \leq \frac{2\varepsilon'}{L_1(4+\sigma_1^2)}$. Then, for any t,*

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] \leq -\tilde{\eta}_t \left(1 - \varepsilon - \varepsilon' - \sigma_1 \mathrm{bias}_t\right) \|\nabla F(\mathbf{w}_t)\|^2 + \widetilde{c}_0 \mathbb{E}\left[\|\boldsymbol{g}_t\|^2 / b_t^2 \mid \mathcal{F}_{t-1}\right],$$

where $\widetilde{c}_0 = \frac{\eta\sigma_0}{2\varepsilon} + \eta^2 \frac{L_0 + \sigma_0 L_1}{2}$ and $\mathrm{bias}_t = \sqrt{\mathbb{E}\left[\|\boldsymbol{g}_t\|^2 / b_t^2 \mid \mathcal{F}_{t-1}\right]}$.

Notice that Lemma 6 only guarantees that the algorithm makes progress on average moving from $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$ when $\sigma_1 \mathrm{bias}_t < 1$, and is essentially vacuous otherwise. To handle this issue, we use the notion of "good times" from (Faw et al., 2022):

Definition 7 (Good times) *A time $t \in [T]$ is "good" if, for fixed parameters $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$ satisfying $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$, $1 - \varepsilon - \varepsilon' - \varepsilon'' - \sigma_1 \mathrm{bias}_t \geq \varepsilon'''$. We denote, for any stopping time τ with respect to $(\mathcal{F}_{s-1})_{s \geq 1}$, $S_{\mathrm{good}}(\tau) = \{1 \leq t < \tau : t \text{ is "good"}\}$ as the set of all such "good" times before τ, and $S_{\mathrm{good}}(\tau)^c = [\tau - 1] \setminus S_{\mathrm{good}}(\tau)$ to be the remaining "bad" times before τ.*

Intuitively, the "good" times are those times when Lemma 6 is non-vacuous. Using Definition 7, we sum the expression Lemma 6 until any stopping time τ to obtain the following more useful form.

Lemma 8 (Descent lemma) *Fix any $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$ such that $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$. For any (L_0, L_1)-function, if we run AdaGrad-Norm with parameters $\eta \leq \frac{2\varepsilon'}{L_1(4+\sigma_1^2)}$ and $b_0^2 > 0$ for T time steps, then, for any stopping time $\tau \in [2, T+1]$ with respect to $(\mathcal{F}_{s-1})_{s \geq 1}$, and any $\widetilde{S}(\tau) \subseteq S_{\mathrm{good}}(\tau)$:*

$$\varepsilon''' \mathbb{E}\left[\sum_{t \in \widetilde{S}(\tau)} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right] \leq F(\mathbf{w}_1) - F^* + 2\widetilde{c}_0 \log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0 \mathbb{E}[\tau - 1]}{\eta\varepsilon'' b_0}\right) + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)} + \mathrm{comp}(\tau),$$

where $\mathrm{comp}(\tau) := \mathbb{E}\left[\sum_{t \in S_{\mathrm{good}}(\tau)^c} (\sigma_1 - (1 - \varepsilon - \varepsilon'))\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t' \in S^{\mathrm{comp}}(\tau)} \varepsilon''' \tilde{\eta}_{t'} \|\nabla F(\mathbf{w}_{t'})\|^2\right]$,

the set $S^{\mathrm{comp}}(\tau) := S_{\mathrm{good}}(\tau) \setminus \widetilde{S}(\tau)$ consists of the "good" times used to compensate for the bad times $S_{\mathrm{good}}(\tau)^c$, and $\widetilde{c}_0 = \frac{\eta\sigma_0}{2\varepsilon} + \eta^2 \frac{L_0 + \sigma_0 L_1}{2}$. In particular, whenever $\sigma_1 \leq 1 - (\varepsilon + \varepsilon')$, then $\mathrm{comp}(\tau) \leq 0$, and when $\sigma_1 \leq 1 - (\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''')$, then additionally $S_{\mathrm{good}}(\tau) = [\tau - 1]$.

5.1. Using the descent lemma when $\sigma_1 < 1$

Let us first analyze Lemma 8 in the simpler setting where $\sigma_1 \leq 1 - (\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''')$. Recall from Lemma 8 that this implies $\mathrm{comp}(\tau) \leq 0$ and we can take $\widetilde{S}(\tau) = S_{\mathrm{good}}(\tau) = [\tau - 1]$. Thus, Lemma 8 loosely becomes $\mathbb{E}\left[\sum_{t<\tau} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \lesssim \log(T)$. At this point, if we choose $\tilde{\eta}_t \approx 1/\sqrt{T}$ and $\tau = T + 1$, then we could conclude that $\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \lesssim \log(T)/\sqrt{T}$. Unfortunately, as we discussed earlier, the first step of our analysis relies on the inequality $\|\mathbf{w}_{t+1} - \mathbf{w}_t\| = \|\eta_t \boldsymbol{g}_t\| \leq 1/L_1$ – a condition which is clearly no longer satisfied when η_t is a fixed constant independent of the gradients. We thus need a different idea to make use of Lemma 8.

We leverage the fact that Lemma 8 holds for *any* stopping time $\tau \in [2, T+1]$ as follows. Suppose there were some stopping time $\tau \in [2, T+1]$ such that:

$$\mathbb{E}\left[\sum_{t<\tau} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \geq \mathbb{E}\left[\tilde{\eta}_\tau\right] \mathbb{E}\left[\sum_{t<\tau} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]. \tag{2}$$

Notice that this inequality would imply that, until τ, we may treat $\tilde{\eta}_t$ and $\|\nabla F(\mathbf{w}_t)\|^2$ as roughly uncorrelated. If (2) were true, we could apply Jensen's inequality and Assumption 4 to obtain:

$$\mathbb{E}\left[\sum_{t<\tau} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \geq \frac{\mathbb{E}\left[\sum_{t<\tau} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]}{\sqrt{\mathbb{E}\left[b_0^2 + \sigma_0^2 T + (1+\sigma_1^2)\mathbb{E}\left[\sum_{t<\tau} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]\right]}}.$$

This, combined with Lemma 8, yields a quadratic inequality in $\sqrt{\mathbb{E}\left[\sum_{t<\tau} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]}$, which can be solved to obtain $\mathbb{E}\left[\sum_{t<\tau} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \lesssim (1+\sigma_1^2)\log(T)^2 + \log(T)\sqrt{b_0^2 + \sigma_0^2 T}$. Thus, if we additionally knew that $\mathbb{E}[\tau] = \Omega(T)$, then a straightforward application of Markov's inequality would imply that, with constant probability, $\min_{t\in[T]} \left\|\nabla F(\mathbf{w}_t)\right\|^2 \lesssim \widetilde{\mathcal{O}}(1/\sqrt{T})$.

It turns out that constructing a time τ (roughly) satisfying (2) is possible – however, there is a tension in simultaneously satisfying this and $\mathbb{E}[\tau] = \Omega(T)$, as the following construction reveals.

Definition 9 (Nice stopping) *Fix any $\delta \in (0, 1]$, and consider the following sequence of random times $\tau_t(\delta)$ defined recursively as follows: let $X_0(\delta) = 1$, and define, for every $t \geq 1$ (denoting $c_{\mathrm{L}} = 2(1+\eta L_1)^2$):*

$$\tau_t(\delta) = \min\left\{t, \min\left\{s \geq 0 : X_s(\delta) = 0\right\}\right\}$$

$$S_t(\delta) = \sum_{s=1}^{\tau_t(\delta)-1} \left\|\boldsymbol{g}_s\right\|^2 + c_{\mathrm{L}} \left\|\nabla F(\mathbf{w}_s)\right\|^2 \quad \textit{and} \quad X_t(\delta) = X_{t-1}(\delta)\mathbb{1}\{S_t(\delta) \leq \mathbb{E}[S_t(\delta)]/\delta\}.$$

Notice that $\tau_1(\delta) = 1$, $S_1(\delta) = 0$, $X_1(\delta) = 1$, and $\tau_2(\delta) = 2$ deterministically. Further, one can show that $S_t(\delta)$, $X_t(\delta)$, and $\tau_{t+1}(\delta)$ are $\mathcal{F}_{t-1}$-measurable for every $t \geq 1$. Intuitively, $\tau_{T+1}(\delta)$ is the first time that the sum of stochastic gradient norms is significantly larger than its expectation, where the expectation is crucially over the random summation range (refer to Remark 24 for a further discussion). The following result shows the utility of this recursive construction:

Lemma 10 (Key properties of nice stopping; Simplified version of Lemma 23) *For any $T \geq 1$ and $\delta \in (0,1]$, let $\tau_{T+1}(\delta)$ be the stopping time from Definition 9. Then, we have the following:*

1. *$\tau_{T+1}(\delta)$ is a stopping time with respect to $(\mathcal{F}_{s-1})_{s\geq 1}$, i.e., $\forall s \geq 1$, $\{s < \tau_{T+1}(\delta)\} \in \mathcal{F}_{s-1}$.*
2. *$\tau_{T+1}(\delta) \in [2, T+1]$, and $\mathbb{E}\left[\tau_{t+1}(\delta)\right] \geq (T+1)(1 - \delta T/2)$.*
3. *For every $s < \tau_{T+1}(\delta)$, denoting $a = b_0^2 + 2\eta^2 L_0^2$ and $b = 1 + \sigma_1^2 + c_{\mathrm{L}}$,*

$$\tilde{\eta}_s \overset{a.s.}{\geq} \frac{\eta}{\sqrt{a + \frac{T\sigma_0^2 + b\mathbb{E}\left[\sum_{\ell < \tau_{T+1}(\delta)} \|\nabla F(\mathbf{w}_\ell)\|^2\right]}{\delta}}},$$

Notice that an immediate consequence of Lemma 10 is that:

$$\mathbb{E}\left[\sum_{t<\tau_{T+1}(\delta)} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right] \geq \frac{\eta \mathbb{E}\left[\sum_{t<\tau_{T+1}(\delta)} \|\nabla F(\mathbf{w}_t)\|^2\right]}{\sqrt{a + \frac{T\sigma_0^2 + b\mathbb{E}\left[\sum_{t<\tau_{T+1}(\delta)} \|\nabla F(\mathbf{w}_t)\|^2\right]}{\delta}}}. \tag{3}$$

When $1/\delta = \mathcal{O}(1)$, then (3) essentially has the desired form (2). Recall that we also needed $\mathbb{E}\left[\tau\right] = \Omega(T)$ to use (2). However, Lemma 10 gives a vacuous lower bound on $\mathbb{E}\left[\tau_{T+1}(\delta)\right]$ when $\delta \geq 2/T$.

Nevertheless, choosing $\delta = \Theta(1/T)$ and solving the resulting quadratic inequality as before, (3) implies $\mathbb{E}\left[\sum_{t<\tau_{T+1}(\delta)} \|\nabla F(\mathbf{w}_t)\|^2\right] \lesssim T\mathrm{poly}\log(T)$. Given that $\mathbb{E}\left[\tau_{T+1}(\delta)\right] \gtrsim T$ in this δ regime, this bound tells us something quite strong – that the sum of gradients before this stopping time scales (roughly) linearly in expectation. This is an *exponential* improvement over the worst-case growth of (L_0, L_1)-smooth functions after T time steps, which is approximately $\exp(L_1\eta T)$. Moreover, this bound implies (via Jensen's inequality) that $\mathbb{E}\left[\tilde{\eta}_{\tau_{T+1}(\delta)}\right] \gtrsim 1/\sqrt{T\mathrm{poly}\log(T)}$! Thus, at least in expectation, the step sizes that we care about for our analysis are essentially scaling as $1/\sqrt{T}$. It turns out that this scaling is crucial to obtain Theorem 3 in the regime of $\sigma_1 < 1$.

5.2. Using the descent lemma when $\sigma_1 \geq 1$

The arguments discussed above heavily relied on being able to take $\mathrm{comp}(\tau) \leq 0$ and $S_{\mathrm{good}}(\tau) = [\tau_{T+1}(\delta) - 1]$, which were trivially true for any stopping time when $\sigma_1 < 1$. However, when $\sigma_1 \geq 1$, then new ideas are needed, since Lemma 6 does not guarantee any meaningful descent inequality for $t \notin S_{\mathrm{good}}(\tau)$. In the context of L_0-smooth optimization, (Faw et al., 2022) showed how to circumvent this issue – indeed, they showed that $\mathrm{comp}(T) \lesssim \mathbb{E}\left[|S_{\mathrm{good}}(T)^c|^2\right]$ and $\mathbb{E}\left[|S_{\mathrm{good}}(T)^c|^2\right] \lesssim \log(T)$. At the core of their proofs for these arguments was the fact that, by L_0-smoothness and properties of (AG-Norm), $|\|\nabla F(\mathbf{w}_t)\| - \|\nabla F(\mathbf{w}_{t'})\|| \lesssim \eta L_0 |t - t'|$.

General (L_0, L_1)-smooth functions clearly violate this inequality. Indeed, $\|\nabla F(\mathbf{w}_t)\|$ can potentially be a multiplicative factor of $\exp(\eta L_1 |t - t'|)$ times larger than $\|\nabla F(\mathbf{w}_{t'})\|$ (for instance, when the $(0, L_1)$-smooth objective is $\exp(L_1 w)$. Thus, even if we could guarantee deterministically that only the first $\mathcal{O}(\log(T))$ time-steps are "bad", the objective function (and also the norm of the gradient) could grow by polynomial factor in T during this interval! In fact, this is exactly the intuition behind our negative result for (AG-Norm) in the "large σ_1" regime (see Lemma 34).

In spite of this, not every (L_0, L_1)-smooth function is an exponential function, as polynomials of constant degree also satisfy (L_0, L_1)-smoothness for constant L_0, L_1 (see Proposition 29). Motivated by this, Definition 4 aims to generalize the inequality $|\|\nabla F(\mathbf{w}_t)\| - \|\nabla F(\mathbf{w}_{t'})\|| \lesssim \eta L_0 |t - t'|$ to allow this difference to have larger polynomial scaling in $t - t'$. Indeed, the constraint of Definition 4 allows us to bound $\mathrm{comp}(\tau)$ as follows:

Lemma 11 *Suppose that $F(\cdot)$ satisfies Definition 4 for some constants $k \geq 2$, $c_k \geq 1$, and $c_k' > 0$. Let $\tau \in [2, T+1]$ be any (possibly random) time. Then, recalling $\mathrm{comp}(\tau)$ and $S^{\mathrm{comp}}(\tau)$ from Lemma 8, there is an explicit construction of $S^{\mathrm{comp}}(\tau)$ (the subset of "good" times used to compensate for $S_{\mathrm{good}}(\tau)^c$) such that, for any $\varepsilon, \varepsilon', \varepsilon''' \in (0,1)$ such that $\varepsilon + \varepsilon' < 1$ and $n_{\mathrm{comp}} = \lceil 4c_k^3(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ / \varepsilon''' \rceil$ (and taking $(x)_+ := \max\{0, x\}$) $\mathrm{comp}(\tau)$ can be bounded as follows:*

$$\mathrm{comp}(\tau) \leq \eta(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ c_k \|\nabla F(\mathbf{w}_1)\| \mathbb{E}\left[|S_{\mathrm{good}}(\tau)^c|\right] + \eta n_{\mathrm{comp}}^{k-1} \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} \left((\sigma_1 - (1-\varepsilon-\varepsilon'))_+ + \frac{\varepsilon''' n_{\mathrm{comp}}}{2c_k^3}\right) \mathbb{E}\left[|S_{\mathrm{good}}(\tau)^c|^k\right].$$

Lemma 11 reveals that, as long as $\mathbb{E}\left[|S_{\mathrm{good}}(\tau_{T+1}(\delta))^c|^k\right]$ can be bounded by $\mathrm{poly}\log(T)$ for any constant $k \geq 1$, then it is still possible to bound $\mathrm{comp}(\tau_{T+1}(\delta))$, even when the function is not L_0-smooth!

Lemma 12 *Let $\tau_{T+1}(\delta) \leq T+1$ be the stopping time with respect to $(\mathcal{F}_{s-1})_{s \geq 1}$ from Definition 9. Recall the set $S_{\mathrm{good}}(\tau_{T+1}(\delta))$ from Definition 21, and denote $S_{\mathrm{good}}(\tau_{T+1}(\delta))^c = [\tau_{T+1}(\delta) - 1] \setminus S_{\mathrm{good}}(\tau_{T+1}(\delta))$. Let $f(T) = e + \frac{e\sigma_0^2(T-1) + e(1+\sigma_1^2+c_{\mathrm{L}})\mathbb{E}\left[\sum_{t<\tau_T(\delta)} \|\nabla F(\mathbf{w}_t)\|^2\right]}{b_0^2 \delta}$. Then, for any $k \geq 1$, the iterates of (AG-Norm) satisfy (under Assumption 4):*

$$\mathbb{E}\left[|S_{\mathrm{good}}(\tau_{T+1}(\delta))^c|^k\right] \leq \left(\frac{(k+1)\sigma_1^2 \log(f(T))}{(1 - (\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon'''))^2}\right)^k.$$

Notice that Lemma 12 does not explicitly require that Definition 4 be satisfied. However, we use this constraint on the objective to easily guarantee that $f(T) = \mathcal{O}(T^{2(k-1)}/\delta)$, and thus that $\mathbb{E}\left[|S_{\mathrm{good}}(\tau_{T+1}(\delta))^c|^k\right] \lesssim \mathrm{poly}\log(T)$. Lemma 31 demonstrates that the bound of $\mathbb{E}\left[|S_{\mathrm{good}}(T)^c|^2\right]$ from (Faw et al., 2022) can be generalized to any moment k. Further, using this generalized result requires bounding only $\mathbb{E}\left[\sum_{t<\tau_T(\delta)} \|\nabla F(\mathbf{w}_t)\|^2\right]$ instead of the sum over the entire time horizon, which might give tighter bounds in some scenarios. With the bounds from Lemmas 11 and 12 in place, it is now clear that a useful descent inequality is still obtainable from Lemma 8 when $\sigma_1 \geq 1$, at least under the added assumption of Definition 4. There is still a (small) problem in translating these results into a convergence result. Indeed, the analogous bound from (3) now becomes:

$$\mathbb{E}\left[\sum_{t \in \widetilde{S}(\tau_{T+1}(\delta))} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right] \geq \frac{\eta \mathbb{E}\left[\sum_{t \in \widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2\right]}{\sqrt{a + \frac{T\sigma_0^2 + b\mathbb{E}\left[\sum_{t<\tau_{T+1}(\delta)} \|\nabla F(\mathbf{w}_t)\|^2\right]}{\delta}}}. \tag{4}$$

Specifically, while the numerator depends on a sum over $\widetilde{S}(\tau_{T+1}(\delta))$, the denominator depends on the sum of these good times, as well as the compensating "good" times $S^{\mathrm{comp}}(\tau_{T+1}(\delta))$ and

the "bad" times before $\tau_{T+1}(\delta)$, $S_{\text{good}}(\tau_{T+1}(\delta))^c$. Faw et al. (2022) dealt with a similar issue by using the fact that, by L_0-smoothness and properties of (AG-Norm), $\|\nabla F(\mathbf{w}_t)\|^2 \lesssim T\log(T/\delta)$ with probability at least $1-\delta$, and $\|\nabla F(\mathbf{w}_t)\|^2 \lesssim T^2$ deterministically. Combining this with their bound $\mathbb{E}\left[|S_{\text{good}}(T)^c|\right] \lesssim \log(T)$, they proved that the sum of "bad" gradients satisfies $\mathbb{E}\left[\sum_{t\in S_{\text{good}}(T)^c}\|\nabla F(\mathbf{w}_t)\|^2\right] \lesssim T\,\text{poly}\log(T)$. However, it is not clear how to prove such a bound in our setting, since $\|\nabla F(\mathbf{w}_t)\|$ can scale as t^{k-1}, which is too large to be useful. Instead, we prove the following *relative* upper bound, which is sufficient for our purposes:

$$\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))^c}\|\nabla F(\mathbf{w}_t)\|^2 \le \mathrm{B}_1 + c_{\mathrm{B}2}\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2,$$

where $\mathbb{E}\left[\mathrm{B}_1\right] \lesssim \text{poly}\log(T)$. As a consequence, (4) becomes:

$$\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2\right] \ge \frac{\eta\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2\right]}{\sqrt{a+\frac{T\sigma_0^2+\mathbb{E}[\mathrm{B}_1]+b\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2\right]}{\delta}}}.$$

Since the numerator and denominator both depend on the same summation, we can apply essentially the same arguments from the $\sigma_1 < 1$ case to obtain our convergence rate for $\sigma_1 \ge 1$ in Theorem 5.

6. The challenges of multiplicative noise for (L_0, L_1)-smooth optimization

Given our positive results for (AG-Norm) from the previous sections, we now turn our focus to algorithms which have been analyzed in prior works on (L_0, L_1)-smooth optimization. Some of the first-studied algorithms for (L_0, L_1)-smooth optimization take the following forms: for parameters $\eta > 0$ and $\gamma \ge 0$:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \frac{\eta \boldsymbol{g}_t}{\gamma + \|\boldsymbol{g}_t\|} \quad \text{and} \quad \mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \frac{\eta \boldsymbol{g}_t}{\max\{\gamma, \|\boldsymbol{g}_t\|\}}. \tag{5}$$

These closely-related updates are referred to as Normalized SGD and Clipped SGD respectively. One motivation for considering these specific updates, at least in the noiseless setting where $\boldsymbol{g}_t = \nabla F(\mathbf{w}_t)$, comes through a comparison with the natural SGD step-size for L_0-smooth non-convex optimization. Indeed, Ghadimi and Lan (2013) show that a constant step-size of $\eta_t = 1/L_0$ yields a $1/T$ rate of convergence to a first-order stationary point. Further, a simple extension of this result (see, e.g., (Bottou et al., 2018) for a proof) is that, under L_0-smoothness and Assumption 4 with $\sigma_0 = 0$ and $\sigma_1 \ge 0$, the step size $\eta_t = 1/L_0(1+\sigma_1^2)$ still achieves the $1/T$ convergence rate. Thus, by analogy, in the (L_0, L_1)-smooth setting, $\eta_t = 1/(L_0 + L_1\|\nabla F(\mathbf{w}_t)\|)$ (and, in the multiplicative noise regime, $\eta_t = 1/(1+\sigma_1^2)(L_0+L_1\|\nabla F(\mathbf{w}_t)\|)$) is a natural candidate step size.

A number of works, including (Zhang et al., 2020b,a; Crawshaw et al., 2022), have proved that (variants of) these algorithms converge whenever the noise of the stochastic gradient satisfies (Bounded-supp). It turns out, however, that these algorithms can diverge under the noise model considered in this paper, (Affine-var). To see this, it is useful to consider a specific stochastic gradient oracle which satisfies Assumptions 3 and 4:

Proposition 13 (A stochastic gradient oracle satisfying Assumption 4) *Fix any $\sigma_0, \sigma_1 \geq 0$, and consider the following stochastic gradient oracle: fix any $\varepsilon \geq 0$, and let, for every $\mathbf{w} \in \mathbb{R}^d$:*

$$\xi_{mult}(\mathbf{w}) = \begin{cases} \left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) & \text{w.p. } \delta = \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2} \\ -\varepsilon & \text{w.p. } 1-\delta \end{cases} \quad \text{and} \quad \xi_{add}(\mathbf{w}) \sim \mathcal{N}(0, \sigma_0^2 I_{d\times d}).$$

We can then take the output of the oracle to be $\boldsymbol{g}(\mathbf{w}) := \xi_{add}(\mathbf{w}) + \xi_{mult}(\mathbf{w})\nabla F(\mathbf{w})$. Then, this construction satisfies Assumptions 3 and 4 with the specified σ_0 and σ_1.

Consider the above oracle with $\sigma_0 = 0$ and $\sigma_1 \gg 1 + \varepsilon$. This oracle outputs stochastic gradients with the same sign as the true gradient for only roughly a $1/\sigma_1^2$ fraction of the times it is queried. The majority of stochastic gradients thus have the *opposite* sign of the true gradient! This turns out to be quite problematic for algorithms of the form (5). Indeed, consider the behavior of (5) when $\|\boldsymbol{g}_t\| \geq \gamma$. In this regime, both algorithms discard the magnitude of the stochastic gradients $\boldsymbol{g}_t$, and use only their sign to perform updates. Since the stochastic gradients $\boldsymbol{g}_t$ of Proposition 13 have the opposite sign of $\nabla F(\mathbf{w}_t)$ for almost all time steps t, one can prove that algorithms of the form (5) do not converge to a stationary point with constant probability under Assumptions 3 and 4, even when the objective function is a 1-dimensional quadratic function (i.e., both smooth and strongly-convex). We give a proof of (a slightly more general version of) this fact in Lemma 35. Similar arguments also imply that (AG-Norm) can diverge on (L_0, L_1)-smooth objectives if η is not tuned with respect to both L_1 and and σ_1, at least in a "large variance" regime (see Lemma 34). For more details, refer to Appendix D.

Acknowledgments

This research is supported in part by NSF Grants 2019844 and 2112471, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

Kengy Barty, Jean-Sébastien Roy, and Cyrille Strugarek. Hilbert-valued perturbed subgradient algorithms. *Mathematics of Operations Research*, 32(3):551–562, 2007.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *arXiv preprint arXiv:2208.11195*, 2022.

Frank E Curtis and Katya Scheinberg. Adaptive stochastic optimization: A framework for analyzing stochastic optimization algorithms. *IEEE Signal Processing Magazine*, 37(5):32–42, 2020.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pages 2658–2667. PMLR, 2020.

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

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 313–355. PMLR, 02–05 Jul 2022.

Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. In *Conference on Learning Theory*, pages 831–875. PMLR, 2017.

Dylan J Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.

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

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jiunn T Hwang. Multiplicative errors-in-variables models with applications to recent data released by the us department of energy. *Journal of the American Statistical Association*, 81(395):680–688, 1986.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

Alejandro Jofré and Philip Thompson. On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Mathematical Programming*, 174(1):253–292, 2019.

Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. In *International Conference on Learning Representations*, 2022.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR, 2019.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24, 2011.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 244–256. Omnipress, 2010.

Arkadi Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Vivak Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on bottou-curtis-nocedal functions. *Mathematical Programming*, 195(1-2):693–734, 2022.

Vivak Patel and Albert S Berahas. Gradient descent in the absence of global lipschitz continuity of the gradients: Convergence, divergence and limitations of its continuous approximation. *arXiv preprint arXiv:2210.02418*, 2022.

Vivak Patel, Shushu Zhang, and Bowen Tian. Global convergence and stability of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:36014–36025, 2022.

BT Polyak and Ya Z Tsypkin. Pseudogradient adaptation and training algorithms. *Automation and remote control*, 34:45–67, 1973.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, 82(3):891–917, 2020.

Robert L Sielken Jr. Stopping times for stochastic approximation procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26(1):67–75, 1973.

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

Donna F Stroup and Henry I Braun. On a new stopping rule for stochastic approximation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 60(4):535–554, 1982.

Bohan Wang, Huishuai Zhang, Zhi-Ming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, Proceedings of Machine Learning Research. PMLR, 12–15 Jul 2023.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21:1–30, 2020.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020b.

Overview of Appendix

A Auxiliary Lemmas **19**
A.1 Useful facts for AdaGrad 19
A.2 Useful facts for (L_0, L_1)-smooth optimization 20
A.3 A note on enforcing $\sigma_1 < 1$ 21

B Proofs for general (L_0, L_1)-smooth functions **22**
B.1 Deriving the descent inequality 22
B.2 Constructing the "nice" stopping time 29
B.3 The key consequence of the nice stopping time construction 33
B.4 Convergence for (L_0, L_1)-smooth functions 35
B.5 A deferred proof for establishing Lemma 6 40

C Proofs for Polynomially-bounded functions for general σ_1 **41**
C.1 The key definition and its properties 42
C.2 Bounding $\mathrm{comp}(\tau)$ from Lemma 8 45
C.3 Bounding the sum of "bad" gradients by the sum of "good" ones 53
C.4 Applying Theorem 26 to polynomially-bounded functions with no restriction on σ_1 57

D Many common algorithms for (L_0, L_1)-smooth optimization can diverge in the presence of multiplicative noise **58**
D.1 Overview of main negative results 59
D.2 Full statement and proof of negative results for (SignSGD-M), (NormSGD), and (ClippedSGD) 61
D.3 Full statement and proof for negative result for (AG-Norm) in the "large σ_1" regime 69

Appendix A. Auxiliary Lemmas

A.1. Useful facts for AdaGrad

Fact 14 *Let $\{a_i\}_{i=1}^{\infty}$ be a sequence of non-negative integers such that $a_1 > 0$. Then, for any T,*

$$\sum_{t\in[T]} \frac{a_t}{\sum_{s=1}^{t} a_s} \le 1 + \log\left(\frac{\sum_{t\in[T]} a_t}{a_1}\right).$$

Proof We proceed via induction. The base case of $T = 1$ holds trivially, with equality. Assuming the hypothesis holds at some time $T \ge 1$, we have that

$$\sum_{t\in[T+1]} \frac{a_t}{\sum_{s=1}^{t} a_s} \le 1 + \log\left(\frac{\sum_{t\in[T]} a_t}{a_1}\right) + \frac{a_{T+1}}{\sum_{s\in[T+1]} a_s}.$$

Now, using the fact that $\exp(x) \le 1/(1-x)$ for any $x < 1$, we have that

$$\frac{a_{T+1}}{\sum_{s\in[T+1]} a_s} = \log\left(\exp\left(\frac{a_{T+1}}{\sum_{s\in[T+1]} a_s}\right)\right) \le \log\left(\frac{1}{1 - \frac{a_{T+1}}{\sum_{s\in[T+1]} a_s}}\right) = \log\left(\frac{\sum_{s\in[T+1]} a_s}{\sum_{s\in[T]} a_s}\right).$$

Combining these two bounds, we conclude that

$$\sum_{t\in[T+1]} \frac{a_t}{\sum_{s=1}^{t} a_s} \le 1 + \log\left(\frac{\sum_{t\in[T]} a_t}{a_1}\right) + \log\left(\frac{\sum_{s\in[T+1]} a_s}{\sum_{s\in[T]} a_s}\right) = 1 + \log\left(\frac{\sum_{t\in[T+1]} a_t}{a_1}\right),$$

so the claim holds also for $T + 1$. Thus, the claim holds for all T by induction. ■

Lemma 15 (Log sum inequality) *The* (AG-Norm) *step-sizes satisfy, for any (possibly random) times $1 \le t_0 \le t_1$ and $s \ge 0$,*

$$\sum_{t=t_0}^{t_1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \le s + \log\left(\frac{b_0^2 + \sum_{t=t_0}^{t_1-s} \|\boldsymbol{g}_t\|^2}{b_0^2}\right)$$

Proof We first note that, by definition of b_t:

$$\sum_{t=t_0}^{t_1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \le \sum_{t=t_0}^{t_1} \frac{\|\boldsymbol{g}_t\|^2}{b_0^2 + \sum_{\ell=t_0}^{t} \|\boldsymbol{g}_\ell\|^2} \le s + \sum_{t=t_0}^{t_1-s} \frac{\|\boldsymbol{g}_t\|^2}{b_0^2 + \sum_{\ell=t_0}^{t} \|\boldsymbol{g}_\ell\|^2}.$$

Thus, applying Fact 14, with $a_1 = b_0^2$ and $a_{\ell+1} = \|\boldsymbol{g}_{t_0+\ell-1}\|^2$ for $\ell \ge 1$, we obtain the claimed inequality. ■

Fact 16 (Bounded Steps) *The iterates $\{\mathbf{w}_s\}_{s=1}^{\infty}$ generated by* (AG-Norm) *satisfy, for every $t \ge 1$,*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \le \eta.$$

Moreover, for any $k \ge 2$ and $t > t'$,

$$\|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1} \le \eta^{k-1}(t - t')^{k-1}.$$

Proof By definition of (AG-Norm),

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\| = \eta_t \|\boldsymbol{g}_t\| = \eta \frac{\|\boldsymbol{g}_t\|}{\sqrt{b_0^2 + \sum_{s=1}^t \|\boldsymbol{g}_s\|^2}} \leq \eta,$$

which establishes the first inequality. To obtain the second, we apply the first, together with Jensen's inequality (noting that $\|\cdot\|^{k-1}$ is convex), to obtain:

$$\begin{aligned}\|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1} = (t-t')^{k-1} \left\| \frac{1}{t-t'} \sum_{s=t'}^{t-1} \mathbf{w}_{s+1} - \mathbf{w}_s \right\|^{k-1} &\leq (t-t')^{k-2} \sum_{s=t'}^{t-1} \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^{k-1} \\ &\leq \eta^{k-1}(t-t')^{k-1},\end{aligned}$$

as claimed. ■

A.2. Useful facts for (L_0, L_1)-smooth optimization

Lemma 17 (Local smoothness bound) *For any function F satisfying Assumption 2, the sequence of iterates $\{\mathbf{w}_s\}_{s=1}^{\infty}$ generated by (AG-Norm) with $\eta \leq 1/L_1$ satisfy*

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \langle \nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{L_0 + L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$$

Proof By (Zhang et al., 2020a, Lemma A.3), we know that, for any function $F(\cdot)$ satisfying Assumption 2, and for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ satisfying $\|\mathbf{w} - \mathbf{w}'\| \leq 1/L_1$,

$$F(\mathbf{w}') \leq F(\mathbf{w}) + \left\langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \right\rangle + \frac{L_0 + L_1 \|\nabla F(\mathbf{w})\|}{2} \left\|\mathbf{w}' - \mathbf{w}\right\|^2.$$

Thus, by choosing $\eta \leq 1/L_1$, the claim is an immediate consequence of Fact 16. ■

Lemma 18 (One-step gradient bound) *For any (L_0, L_1)-smooth function $F(\cdot)$, assuming that $\eta \leq 1/L_1$, the gradient $\|\nabla F(\mathbf{w}_t)\|^2$ evaluated at the iterate of (AG-Norm) at time t satisfies:*

$$\|\nabla F(\mathbf{w}_t)\|^2 \leq 2\eta^2 L_0^2 + 2(1 + \eta L_1)^2 \|\nabla F(\mathbf{w}_{t-1})\|^2.$$

Proof Since $\eta \leq 1/L_1$, $\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \leq 1/L_1$ by Fact 16. Thus, we may apply Assumption 2 to obtain

$$\begin{aligned}\|\nabla F(\mathbf{w}_t)\| &\leq \|\nabla F(\mathbf{w}_{t-1})\| + \|\nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t-1})\| \\ &\leq \|\nabla F(\mathbf{w}_{t-1})\| + (L_0 + L_1 \|\nabla F(\mathbf{w}_{t-1})\|)\eta_{t-1} \|\boldsymbol{g}_{t-1}\|,\end{aligned}$$

from which we conclude that:

$$\|\nabla F(\mathbf{w}_t)\|^2 \leq 2\eta^2 L_0^2 + 2(1 + \eta L_1)^2 \|\nabla F(\mathbf{w}_{t-1})\|^2.$$

■

Proposition 1 *A function satisfying (L_0, L_1)-smoothness according to* (Generalized-smooth) *is also $(2L_0, (e-1)L_1)$-smooth according to Assumption 2. If $F(\cdot)$ is twice continuously differentiable and (L_0, L_1)-smooth according to Assumption 2, then it is also (L_0, L_1)-smooth according to* (Generalized-smooth)*.*

Proof The proof of the first statement is from (Zhang et al., 2020a, Corollary A.4). The proof of the second statement closely follows the analogous proof for L_0-smooth functions from (Nesterov, 2003, Lemma 1.2.2). We give a proof of this claim for completeness.

Consider any $\mathbf{x}, \mathbf{s} \in \mathbb{R}^d$ such that $0 < \|\mathbf{s}\| \leq 1/L_1$, and let $\alpha \in (0, 1]$. Then, by Assumption 2,

$$\left\| \frac{\nabla F(\mathbf{x} + \alpha \mathbf{s}) - \nabla F(\mathbf{x})}{\alpha} \right\| \leq (L_0 + L_1 \|\nabla F(\mathbf{x})\|) \|\mathbf{s}\| .$$

Therefore, we have the following:

$$\begin{aligned}
&\lim_{\alpha \to 0} \left\| \frac{\nabla F(\mathbf{x} + \alpha \mathbf{s}) - \nabla F(\mathbf{x})}{\alpha} \right\| \\
&= \left\| \lim_{\alpha \to 0} \frac{\nabla F(\mathbf{x} + \alpha \mathbf{s}) - \nabla F(\mathbf{x})}{\alpha} \right\| && \text{by continuity of } \|\cdot\| \text{ and twice differentiability of } F(\cdot) \\
&= \left\| \nabla^2 F(\mathbf{x}) \cdot \mathbf{s} \right\| && \text{by definition of directional derivative}
\end{aligned}$$

Hence, by the limit inequality theorem, we have that, for any $0 < \|\mathbf{s}\| \leq 1/L_1$,

$$\frac{\left\| \nabla^2 F(\mathbf{x}) \cdot \mathbf{s} \right\|}{\|\mathbf{s}\|} \leq L_0 + L_1 \|\nabla F(\mathbf{x})\| .$$

In particular, by taking the supremum over all such $\mathbf{s}$, we conclude that

$$\left\| \nabla^2 F(\mathbf{x}) \right\| = \left\| \nabla^2 F(\mathbf{x})^\top \right\| \leq L_0 + L_1 \|\nabla F(\mathbf{x})\| ,$$

as claimed, where the first equality follows by observing that $\nabla^2 F(\mathbf{x}) \nabla^2 F(\mathbf{x})^\top$ and $\nabla^2 F(\mathbf{x})^\top \nabla^2 F(\mathbf{x})$ have the same non-zero eigenvalues (since all entries of $\nabla^2 F(\mathbf{x})$ are real, and by appealing to the singular value decomposition), which implies that $\nabla^2 F(x)$ and $\nabla^2 F(x)^\top$ have the same spectral norm. ■

A.3. A note on enforcing $\sigma_1 < 1$

Fact 19 (Reducing σ_1 through mini-batching) *Suppose that the stochastic gradient oracle satisfies Assumptions 3 and 4 for some $\sigma_0 \geq 0$ and $\sigma_1 \geq 1$. Then, assuming this oracle returns independent stochastic gradients each time $\boldsymbol{g}(\mathbf{w})$ is sampled, one can construct, for any $\varepsilon \in (0, 1)$, a new stochastic gradient oracle from this one through mini-batching which satisfies Assumptions 3 and 4 with $\widetilde{\sigma_0} \leq \sigma_0$ and $\widetilde{\sigma_1} = 1 - \varepsilon$, and where each call to the new gradient requires only $B = \lceil \sigma_1^2 / (1-\varepsilon)^2 \rceil$ calls to the old one.*

Proof Fix any $\varepsilon \in (0,1)$ and $\mathbf{w} \in \mathbb{R}^d$. Let $B = \lceil \sigma_1^2/(1-\varepsilon)^2 \rceil$, and let $\{\boldsymbol{g}_j(\mathbf{w})\}_{j\in[B]}$ be a set of B independent stochastic gradients corresponding to $\nabla F(\mathbf{w})$ from an oracle satsifying Assumptions 3 and 4 with $\sigma_0 \geq 0$ and $\sigma_1 \geq 1$. Then, we take the response of the new oracle as:

$$\widetilde{\boldsymbol{g}}(\mathbf{w}) := \frac{1}{B}\sum_{j\in[B]} \boldsymbol{g}_j(\mathbf{w}).$$

Now, since $\mathbb{E}\left[\boldsymbol{g}_j(\mathbf{w})\right] = \nabla F(\mathbf{w})$ and applying linearity of expectation, $\mathbb{E}\left[\widetilde{\boldsymbol{g}}(\mathbf{w})\right] = \nabla F(\mathbf{w})$. Further, notice that:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}(\mathbf{w}) - \nabla F(\mathbf{w})\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{j\in[B]} \boldsymbol{g}_j(\mathbf{w}) - \nabla F(\mathbf{w})\right\|^2\right] \\
&= \frac{1}{B^2}\sum_{j\in[B]} \mathbb{E}\left[\left\|\boldsymbol{g}_j(\mathbf{w}) - \nabla F(\mathbf{w})\right\|^2\right] \\
&\quad + \frac{2}{B^2}\sum_{B\geq j>j'\geq 1} \mathbb{E}\left[\left\langle \boldsymbol{g}_j(\mathbf{w}) - \nabla F(\mathbf{w}), \boldsymbol{g}_{j'}(\mathbf{w}) - \nabla F(\mathbf{w})\right\rangle\right] \\
&\leq \frac{1}{B^2}\sum_{j\in[B]} \sigma_0^2 + \sigma_1^2\left\|\nabla F(\mathbf{w})\right\|^2 \\
&\leq \frac{(1-\varepsilon)^2\sigma_0^2}{\sigma_1^2} + (1-\varepsilon)^2\left\|\nabla F(\mathbf{w})\right\|^2,
\end{aligned}$$

where the first inequality follows by Assumption 3 and since $\boldsymbol{g}_j(\mathbf{w})$ and $\boldsymbol{g}_{j'}(\mathbf{w})$ are independent. The second inequality follows by Assumption 4 and our choice of $B \geq \sigma_1^2/(1-\varepsilon)^2$. Thus, Assumption 4 is satisfied with $\widetilde{\sigma_0}^2 = (1-\varepsilon)^2\sigma_0^2/\sigma_1^2 \leq \sigma_0^2$ and $\widetilde{\sigma_1}^2 = (1-\varepsilon)^2$. ■

Appendix B. Proofs for general (L_0, L_1)-smooth functions

B.1. Deriving the descent inequality

The following inequality serves as the first step in analyzing the convergence of (AG-Norm).

Lemma 6 *Fix any $\varepsilon, \varepsilon' \in (0,1)$. Suppose that $\eta \leq \frac{2\varepsilon'}{L_1(4+\sigma_1^2)}$. Then, for any t,*

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] \leq -\tilde{\eta}_t\left(1 - \varepsilon - \varepsilon' - \sigma_1\mathrm{bias}_t\right)\left\|\nabla F(\mathbf{w}_t)\right\|^2 + \widetilde{c}_0\mathbb{E}\left[\|\boldsymbol{g}_t\|^2/b_t^2 \mid \mathcal{F}_{t-1}\right],$$

where $\widetilde{c}_0 = \frac{\eta\sigma_0}{2\varepsilon} + \eta^2\frac{L_0+\sigma_0 L_1}{2}$ and $\mathrm{bias}_t = \sqrt{\mathbb{E}\left[\|\boldsymbol{g}_t\|^2/b_t^2 \mid \mathcal{F}_{t-1}\right]}$.

Proof An immediate consequence of Lemma 17 and (Faw et al., 2022, Lemma 5) is that, as long as $\eta \leq 1/L_1$,

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] &\leq -\tilde{\eta}_t\left(1 - \varepsilon - \sigma_1\mathrm{bias}_t\right)\left\|\nabla F(\mathbf{w}_t)\right\|^2 + c_0\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \\
&\quad + \frac{L_1\left\|\nabla F(\mathbf{w}_t)\right\|}{2}\mathbb{E}\left[\eta_t^2\left\|\boldsymbol{g}_t\right\|^2 \mid \mathcal{F}_{t-1}\right], \qquad (6)
\end{aligned}$$

where

$$c_0 = \frac{\eta\sigma_0}{2\varepsilon} + \frac{\eta^2 L_0}{2} \quad \text{and} \quad \text{bias}_t = \sqrt{\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2 \left(\|\widetilde{\nabla}_t\| + \|\boldsymbol{g}_t\|\right)^2}{b_t^2(\widetilde{b}_t + b_t)^2} \mid \mathcal{F}_{t-1}\right]}.$$

We provide a proof of this inequality in Lemma 27[3].

Now, let's focus on bounding the final term above. We start by rewriting it as follows: Let us take $\mathcal{E}_{\sigma_0} = \{\|\nabla F(\mathbf{w}_t)\| > \sigma_0\}$. Then, we can decompose the final term (trivially) as

$$\frac{L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] = \frac{L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] (\mathbb{1}\{\mathcal{E}_{\sigma_0}^c\} + \mathbb{1}\{\mathcal{E}_{\sigma_0}\}).$$

Now, whenever $\mathcal{E}_{\sigma_0}$ is false, then this expression is easy to bound, since

$$\frac{L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{\mathcal{E}_{\sigma_0}^c\} \leq \frac{L_1\sigma_0}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right].$$

Notice that this term can be absorbed into the second term in (6). The case when $\mathcal{E}_{\sigma_0}$ is true requires slightly more care. However, we can deal with this case by adding and subtracting $\tilde{\eta}_t^2$, and using the bound (1):

$$\begin{aligned}
\frac{L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{\mathcal{E}_{\sigma_0}\} &= \frac{L_1}{2}\tilde{\eta}_t^2 \|\nabla F(\mathbf{w}_t)\| \mathbb{E}\left[\|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{\mathcal{E}_{\sigma_0}\} \\
&\quad + \frac{L_1}{2} \|\nabla F(\mathbf{w}_t)\| \mathbb{E}\left[(\eta_t^2 - \tilde{\eta}_t^2) \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{\mathcal{E}_{\sigma_0}\} \\
&\leq \frac{L_1(2+\sigma_1^2)}{2}\tilde{\eta}_t^2 \|\nabla F(\mathbf{w}_t)\|^3 \\
&\quad + \frac{L_1}{2} \|\nabla F(\mathbf{w}_t)\| \mathbb{E}\left[(\eta_t^2 - \tilde{\eta}_t^2) \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{\mathcal{E}_{\sigma_0}\} \\
&\leq \frac{\eta L_1(2+\sigma_1^2)}{2}\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 \\
&\quad + \frac{L_1}{2} \|\nabla F(\mathbf{w}_t)\| \mathbb{E}\left[(\eta_t^2 - \tilde{\eta}_t^2) \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{\mathcal{E}_{\sigma_0}\}.
\end{aligned}$$

3. A careful reader may notice that the inequality in Lemma 27 is actually slightly smaller than the one from (Faw et al., 2022, Lemma 5), since the dependence on constants is strictly better.

Notice that the first term above can be absorbed into the first term in (6), assuming η is sufficiently small. For the remaining term, we begin by noticing that

$$\begin{aligned}\frac{\eta_t^2-\tilde{\eta}_t^2}{\eta^2}\mathbb{1}\{\mathcal{E}_{\sigma_0}\} &= \frac{\mathbb{1}\{\mathcal{E}_{\sigma_0}\}}{b_{t-1}^2+\|\boldsymbol{g}_t\|^2}-\frac{\mathbb{1}\{\mathcal{E}_{\sigma_0}\}}{b_{t-1}^2+\|\widetilde{\nabla}_t\|^2}\\ &= \frac{(\|\widetilde{\nabla}_t\|^2-\|\boldsymbol{g}_t\|^2)\mathbb{1}\{\mathcal{E}_{\sigma_0}\}}{(b_{t-1}^2+\|\boldsymbol{g}_t\|^2)(b_{t-1}^2+\|\widetilde{\nabla}_t\|^2)}\\ &\leq \frac{\|\widetilde{\nabla}_t\|^2\mathbb{1}\{\mathcal{E}_{\sigma_0}\}}{(b_{t-1}^2+\|\boldsymbol{g}_t\|^2)(b_{t-1}^2+\|\widetilde{\nabla}_t\|^2)}\\ &\leq \frac{2\|\nabla F(\mathbf{w}_t)\|^2}{(b_{t-1}^2+\|\boldsymbol{g}_t\|^2)(b_{t-1}^2+\|\widetilde{\nabla}_t\|^2)}\\ &\leq \frac{2\|\nabla F(\mathbf{w}_t)\|}{(b_{t-1}^2+\|\boldsymbol{g}_t\|^2)\sqrt{b_{t-1}^2+\|\widetilde{\nabla}_t\|^2}},\end{aligned}$$

which implies that

$$\begin{aligned}\frac{L_1(\eta_t^2-\tilde{\eta}_t^2)}{2}\|\nabla F(\mathbf{w}_t)\|\|\boldsymbol{g}_t\|^2\mathbb{1}\{\mathcal{E}_{\sigma_0}\} &\leq \eta L_1\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\\ &\leq \eta L_1\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2.\end{aligned}$$

Therefore, collecting these results and choosing $\eta < {}^{2\varepsilon'}/_{L_1(4+\sigma_1^2)}$, we have that

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{w}_{t+1})-F(\mathbf{w}_t)\mid\mathcal{F}_{t-1}\right] \leq -\tilde{\eta}_t\left(1-\varepsilon-\sigma_1\mathrm{bias}_t-\frac{\eta L_1(4+\sigma_1^2)}{2}\right)\|\nabla F(\mathbf{w}_t)\|^2+\widetilde{c}_0\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mid\mathcal{F}_{t-1}\right]\\ &\leq -\tilde{\eta}_t\left(1-\varepsilon-\varepsilon'-\sigma_1\mathrm{bias}_t\right)\|\nabla F(\mathbf{w}_t)\|^2+\widetilde{c}_0\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mid\mathcal{F}_{t-1}\right],\end{aligned}$$

where $\widetilde{c}_0 = c_0+\eta^2L_1\sigma_0$. ■

We use Lemma 6 by summing the expression until a carefully-chosen stopping time. To make use of this bound, we begin by showing that the second additive term can (essentially) be absorbed into the first term.

Lemma 20 *Fix any $\varepsilon''\in(0,1)$, and let $2\leq\tau\leq T+1$ be any stopping time with respect to $(\mathcal{F}_{s-1})_{s\geq1}$. Then, we have that*

$$\widetilde{c}_0\mathbb{E}\left[\sum_{t=1}^{\tau-1}\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mid\mathcal{F}_{t-1}\right]\leq\varepsilon''\mathbb{E}\left[\sum_{t=1}^{\tau-1}\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2\right]+2\widetilde{c}_0\log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0\mathbb{E}\left[\tau-1\right]}{\eta\varepsilon''b_0}\right)+\frac{2\eta\varepsilon''\sigma_0}{2+\sigma_1^2}.$$

Proof Let us define, for a parameter λ to be determined,

$$\tau_{\mathrm{step}}(\lambda)=\min\left\{T+1,\min\left\{t\geq1:\eta_t\leq\lambda\right\}\right\}.$$

By construction, $\tau_{\text{step}}(\lambda)$ is the first time when the step size η_t is smaller than some threshold λ (or $T+1$ in the case that η_t remains larger than λ for every $t \in [T]$). Observe that $\eta_t > \lambda$ is equivalent to $b_t < \eta/\lambda$. Thus, we divide our analysis into two phases: times before $\tau_{\text{step}}(\lambda)$, and those after. For the earlier times, since we have $b_{\tau_{\text{step}}(\lambda)-1} < \eta/\lambda$, we can bound these using Lemma 15. We use the fact that $\eta_t \le \lambda$ together with (1) to handle the remaining terms.

More specifically, for any t, we can decompose

$$\widetilde{c}_0 \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} = \widetilde{c}_0 \frac{\|\boldsymbol{g}_t\|^2}{b_t^2}(\mathbb{1}\{t < \tau_{\text{step}}(\lambda)\} + \mathbb{1}\{t \ge \tau_{\text{step}}(\lambda)\}).$$

Now, note that, by definition of $\tau_{\text{step}}(\lambda)$, $\eta_{\tau_{\text{step}}(\lambda)-1} > \lambda$, i.e., $b_{\tau_{\text{step}}(\lambda)-1} < \eta/\lambda$. Hence, by Lemma 15,

$$\sum_{t=1}^{\tau-1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mathbb{1}\{t < \tau_{\text{step}}(\lambda)\} \le \sum_{t < \tau_{\text{step}}(\lambda)} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \le 2\log(b_{\tau_{\text{step}}(\lambda)-1}/b_0) \le 2\log\left(\frac{\eta}{\lambda b_0}\right).$$

In the other case, for any fixed $t \in [T]$, we have that

$$\begin{aligned}
\eta_t^2 \|\boldsymbol{g}_t\|^2 \mathbb{1}\{t \ge \tau_{\text{step}}(\lambda)\} &\le \lambda \eta_t \|\boldsymbol{g}_t\|^2 \\
&= \lambda(\eta_t - \tilde{\eta}_t)\|\boldsymbol{g}_t\|^2 + \lambda \tilde{\eta}_t \|\boldsymbol{g}_t\|^2 \\
&\le \eta\lambda \frac{\|\widetilde{\nabla}_t\|^2}{b_t\sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2}(b_t + \sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2})} \|\boldsymbol{g}_t\|^2 + \lambda \tilde{\eta}_t \|\boldsymbol{g}_t\|^2 \\
&\le \lambda \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 + \lambda\eta\sigma_0 + \lambda \tilde{\eta}_t \|\boldsymbol{g}_t\|^2,
\end{aligned}$$

where, in the first inequality, we used the fact that $\eta_t \le \lambda$ for every $t \ge \tau_{\text{step}}(\lambda)$, and in the second, we used the fact that

$$\begin{aligned}
\frac{\eta_t - \tilde{\eta}_t}{\eta} = \frac{1}{\sqrt{b_{t-1}^2 + \|\boldsymbol{g}_t\|^2}} - \frac{1}{\sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2}} &= \frac{\|\widetilde{\nabla}_t\|^2 - \|\boldsymbol{g}_t\|^2}{b_t\sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2}(b_t + \sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2})} \\
&\le \frac{\|\widetilde{\nabla}_t\|^2}{b_t^2\sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2}},
\end{aligned}$$

and in the third, we used the fact that $\|\widetilde{\nabla}_t\|^2 = \sigma_0^2 + \|\nabla F(\mathbf{w}_t)\|^2$. Then, noting that

$$\lambda\tilde{\eta}_t \mathbb{E}\left[\|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \le \lambda(1+\sigma_1^2)\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 + \lambda\tilde{\eta}_t\sigma_0^2 \le \lambda(1+\sigma_1^2)\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 + \lambda\eta\sigma_0,$$

we have that

$$\mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mathbb{1}\{t \ge \tau_{\text{step}}(\lambda)\} \mid \mathcal{F}_{t-1}\right] \le 2\lambda\eta\sigma_0 + \lambda(2+\sigma_1^2)\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2.$$

Combining these bounds, we obtain:

$$\widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \le \tilde{\eta}_t \frac{\lambda\widetilde{c}_0(2+\sigma_1^2)}{\eta^2} \|\nabla F(\mathbf{w}_t)\|^2 + \widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mathbb{1}\{t < \tau_{\text{step}}(\lambda)\} \mid \mathcal{F}_{t-1}\right] + \frac{2\widetilde{c}_0\lambda\sigma_0}{\eta}$$

Thus, if we choose $\lambda = \frac{\varepsilon'' \eta^2}{(2+\sigma_1^2)\widetilde{c}_0(\tau-1)}$, then we obtain

$$\widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \leq \varepsilon'' \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 + \widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mathbb{1}\{t < \tau_{\text{step}}(\lambda)\} \mid \mathcal{F}_{t-1}\right] + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)(\tau-1)}.$$

Now, summing over $t \in [\tau - 1]$, and using the fact that $\{t < \tau\} \in \mathcal{F}_{t-1}$ by assumption on τ, we have:

$$\begin{aligned}
\widetilde{c}_0 \mathbb{E}\left[\sum_{t=1}^{\tau-1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\right] &= \widetilde{c}_0 \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mathbb{1}\{t < \tau\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \mathbb{1}\{t < \tau\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{\tau-1} \varepsilon'' \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right] + \widetilde{c}_0 \mathbb{E}\left[\sum_{t=1}^{\tau-1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mathbb{1}\{t < \tau_{\text{step}}(\lambda)\} + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)(\tau-1)}\right].
\end{aligned}$$

Focusing on the last term in the above inequality, and recalling that (deterministically) $\tau > 1$ by assumption, we may apply the above bounds together with Jensen's inequality to obtain:

$$\begin{aligned}
\widetilde{c}_0 \mathbb{E}\left[\sum_{t=1}^{\tau-1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mathbb{1}\{t < \tau_{\text{step}}(\lambda)\} + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)(\tau-1)}\right] &\leq 2\widetilde{c}_0 \mathbb{E}\left[\log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0(\tau-1)}{\eta\varepsilon'' b_0}\right)\right] + \frac{2\eta\varepsilon''\sigma_0}{2+\sigma_1^2} \\
&\leq 2\widetilde{c}_0 \log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0 \mathbb{E}\left[\tau - 1\right]}{\eta\varepsilon'' b_0}\right) + \frac{2\eta\varepsilon''\sigma_0}{2+\sigma_1^2}.
\end{aligned}$$

Combining these bounds yields the claimed inequality. ■

In the following, we restate Definition 7 with an equivalent characterization that is sometimes more convenient for our analysis.

Definition 21 (Good times (extended version of Definition 7)) *A time $t \in [T]$ is "good" if, for fixed parameters $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$, satisfying $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$*

$$1 - \varepsilon - \varepsilon' - \varepsilon'' - \sigma_1 \text{bias}_t \geq \varepsilon''' \quad \textit{or, equivalently,} \quad \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \leq \frac{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}{\sigma_1^2}.$$

We take, for any stopping time τ with respect to $(\mathcal{F}_{s-1})_{s\geq 1}$, the set $S_{\text{good}}(\tau) = \{1 \leq t < \tau : t \text{ is "good"}\}$ to be the "good" times before τ, and $S_{\text{good}}(\tau)^c = [\tau - 1] \setminus S_{\text{good}}(\tau)$ to be the remaining "bad" times before τ.

We will now use the notion of "good" and "bad" times from Definition 21 to understand how the bias_t term affects Lemma 6.

Lemma 22 (Bounds for "good" and "bad" times) *Consider the same setting as Lemmas 6 and 20. Let $\tau \in [T+1]$ be any stopping time with respect to $(\mathcal{F}_{s-1})_{s\geq 1}$. Then, for any $t \in S_{\text{good}}(\tau)$,*

$$(\varepsilon'' + \varepsilon''')\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 \leq \mathbb{E}\left[F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \mid \mathcal{F}_{t-1}\right] + \widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right].$$

For every other time $t \notin S_{\mathrm{good}}(\tau)$, we have that

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] \leq \tilde{\eta}_t \left(\sigma_1 - \left(1 - (\varepsilon + \varepsilon')\right)\right) \|\nabla F(\mathbf{w}_t)\|^2 + \widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right].$$

In particular, whenever $\sigma_1 \leq (1 - \varepsilon - \varepsilon')$, then we have the following bound for each $t \notin S_{\mathrm{good}}(\tau)$.

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] \leq \widetilde{c}_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \leq \widetilde{c}_0.$$

Proof We note that, by construction, $\{t < \tau\}, \{t \in S_{\mathrm{good}}(\tau)\} \in \mathcal{F}_{t-1}$, since τ is a stopping time with respect to $(\mathcal{F}_{s-1})_{s \geq 1}$ and $\mathbb{E}\left[\|\boldsymbol{g}_t\|^2 / b_t^2 \mid \mathcal{F}_{t-1}\right]$ is $\mathcal{F}_{t-1}$-measurable. Since the inequalities we wish to prove are in expectation conditioned on $\mathcal{F}_{t-1}$, the condition that a time t is "good" or "bad" is (effectively) deterministic.

The proof of the first inequality is an immediate consequence of Lemmas 6 and 20 and Definition 21. The second follows immediately from Lemma 6, noting that $\mathrm{bias}_t = \sqrt{\mathbb{E}\left[\|\boldsymbol{g}_t\|^2 / b_t^2 \mid \mathcal{F}_{t-1}\right]} \leq 1$. The final follows from the second, noting in this case that $\sigma_1 - (1 - (\varepsilon + \varepsilon')) \leq 0$. ■

We now combine the results from Lemmas 6, 20 and 22 to obtain our main descent lemma. This gives us a bound on $\mathbb{E}\left[\sum_{t \in \widetilde{S}(\tau)} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right]$ in terms of negligible terms and a "compensation" term, $\mathrm{comp}(\tau)$, where the summation is taken over a random subset of the "good" times.

Lemma 8 (Descent lemma) *Fix any $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$ such that $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$. For any (L_0, L_1)-function, if we run AdaGrad-Norm with parameters $\eta \leq \frac{2\varepsilon'}{L_1(4+\sigma_1^2)}$ and $b_0^2 > 0$ for T time steps, then, for any stopping time $\tau \in [2, T+1]$ with respect to $(\mathcal{F}_{s-1})_{s \geq 1}$, and any $\widetilde{S}(\tau) \subseteq S_{\mathrm{good}}(\tau)$:*

$$\varepsilon''' \mathbb{E}\left[\sum_{t \in \widetilde{S}(\tau)} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right] \leq F(\mathbf{w}_1) - F^* + 2\widetilde{c}_0 \log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0 \mathbb{E}[\tau - 1]}{\eta \varepsilon'' b_0}\right) + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)} + \mathrm{comp}(\tau),$$

where $\mathrm{comp}(\tau) := \mathbb{E}\left[\sum_{t \in S_{\mathrm{good}}(\tau)^c} (\sigma_1 - (1 - \varepsilon - \varepsilon')) \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t' \in S^{\mathrm{comp}}(\tau)} \varepsilon''' \tilde{\eta}_{t'} \|\nabla F(\mathbf{w}_{t'})\|^2\right]$,
the set $S^{\mathrm{comp}}(\tau) := S_{\mathrm{good}}(\tau) \setminus \widetilde{S}(\tau)$ consists of the "good" times used to compensate for the bad times $S_{\mathrm{good}}(\tau)^c$, and $\widetilde{c}_0 = \frac{\eta\sigma_0}{2\varepsilon} + \eta^2 \frac{L_0 + \sigma_0 L_1}{2}$. In particular, whenever $\sigma_1 \leq 1 - (\varepsilon + \varepsilon')$, then $\mathrm{comp}(\tau) \leq 0$, and when $\sigma_1 \leq 1 - (\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''')$, then additionally $S_{\mathrm{good}}(\tau) = [\tau - 1]$.

Proof The proof follows straightforwardly by combining the inequalities from Lemma 22, together with noting that, since τ is a stopping time with respect to $(\mathcal{F}_{s-1})_{s \geq 1}$, $\{s < \tau\} \in \mathcal{F}_{s-1}$. Indeed, since $[\tau - 1] = S_{\mathrm{good}}(\tau) \cup S_{\mathrm{good}}(\tau)^c$, we may apply the tower rule and linearity of expectation to

conclude that

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_\tau) - F(\mathbf{w}_1)\right] &= \mathbb{E}\left[\sum_{t<\tau} F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)\right] \\ &= \sum_{t\in[T]} \mathbb{E}\left[\mathbb{E}\left[(F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t))\mathbb{1}\{t<\tau\} \mid \mathcal{F}_{t-1}\right]\right] \\ &= \sum_{t\in[T]} \mathbb{E}\left[\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right]\mathbb{1}\{t<\tau\}\right] \\ &= \sum_{t\in[T]} \mathbb{E}\left[\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right]\mathbb{1}\{t\in S_{\text{good}}(\tau)\}\right] \\ &\quad + \sum_{t\in[T]} \mathbb{E}\left[\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right]\mathbb{1}\{t\in S_{\text{good}}(\tau)^c\}\right]\end{aligned}$$

Now, we may use the first and second inequalities in Lemma 22 to bound the sum over "good" and "bad" times, respectively, and, collecting terms, we obtain

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_\tau) - F(\mathbf{w}_1)\right] \leq -(\varepsilon'' + \varepsilon''')\mathbb{E}&\left[\sum_{t\in S_{\text{good}}(\tau)} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \\ &+ \mathbb{E}\left[\sum_{t\in S_{\text{good}}(\tau)^c} \left(\sigma_1 - \left(1-(\varepsilon+\varepsilon')\right)\right)\tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \\ &+ \widetilde{c}_0\mathbb{E}\left[\sum_{t\in[\tau-1]} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\right]\end{aligned}$$

Thus, applying Lemma 20 to bound the final term above, and using the fact that $\mathbb{E}\left[F(\mathbf{w}_1) - F(\mathbf{w}_\tau)\right] \leq F(\mathbf{w}_1) - F^*$ by Assumption 1, we obtain:

$$\begin{aligned}\varepsilon'''\mathbb{E}\left[\sum_{t\in S_{\text{good}}(\tau)} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \leq F(\mathbf{w}_1) - F^* + 2\widetilde{c}_0 \log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0\mathbb{E}\left[\tau-1\right]}{\eta\varepsilon''b_0}\right) + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)} \\ + \mathbb{E}\left[\sum_{t\in S_{\text{good}}(\tau)^c} \left(\sigma_1 - \left(1-\varepsilon-\varepsilon'\right)\right)\tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right).\end{aligned}$$

Thus, for any $\widetilde{S}(\tau) \subset S_{\text{good}}(\tau)$, we can subtract $\varepsilon'''\mathbb{E}\left[\sum_{t\in S_{\text{good}}(\tau)\setminus\widetilde{S}(\tau)} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]$ from both sides of the above inequality to obtain the first claimed inequality.

The second follows immediately by noting in this case that $\sigma_1 - (1-(\varepsilon+\varepsilon')) \leq 0$. The third follows immediately from the second, recalling that, whenever $\sigma_1 \leq (1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))$, then $S_{\text{good}}(\tau) = [\tau-1]$ by Definition 21. ■

B.2. Constructing the "nice" stopping time

Let us recall the definition of $\tau_{T+1}(\delta)$, the "nice" stopping time:

Definition 9 (Nice stopping) *Fix any $\delta \in (0,1]$, and consider the following sequence of random times $\tau_t(\delta)$ defined recursively as follows: let $X_0(\delta) = 1$, and define, for every $t \geq 1$ (denoting $c_{\mathrm{L}} = 2(1+\eta L_1)^2$):*

$$\tau_t(\delta) = \min\left\{t, \min\left\{s \geq 0 : X_s(\delta) = 0\right\}\right\}$$
$$S_t(\delta) = \sum_{s=1}^{\tau_t(\delta)-1} \|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}} \|\nabla F(\mathbf{w}_s)\|^2 \quad \text{and} \quad X_t(\delta) = X_{t-1}(\delta)\mathbb{1}\{S_t(\delta) \leq {}^{\mathbb{E}[S_t(\delta)]}/_{\delta}\}.$$

Here, we show that these random variables are well-defined, and enumerate the crucial properties that they satisfy.

Lemma 23 (Nice stopping; Full version of Lemma 10) *For any $\delta \in (0,1]$ and $t \geq 1$, let $\tau_t(\delta)$, $S_t(\delta)$, and $X_t(\delta)$ be recursively-defined random variables from Definition 9. Then, we have that, for all $t \geq 1$,*

1. *$\tau_t(\delta)$ is $\mathcal{F}_{t-2}$-measurable, and $S_t(\delta), X_t(\delta)$ are each $\mathcal{F}_{t-1}$-measurable (where we take $\mathcal{F}_0 = \mathcal{F}_{-1}$ to be the trivial σ-algebra).*
2. *$\tau_t(\delta)$ is a stopping time with respect to $(\mathcal{F}_{s-1})_{s\geq 1}$, i.e., for all $s \geq 0$ $\{s < \tau_t(\delta)\} \in \mathcal{F}_{s-1}$.*
3. *For all $t \geq 1$, $\tau_{t+1}(\delta) \geq \tau_t(\delta)$, $S_{t+1}(\delta) \geq S_t(\delta)$, and $X_{t+1}(\delta) \leq X_t(\delta)$.*
4. *$\mathbb{E}\left[S_t(\delta)\right] \leq \mathbb{E}\left[\sum_{s<\tau_t(\delta)} \sigma_0^2 + (1+\sigma_1^2+c_{\mathrm{L}})\|\nabla F(\mathbf{w}_s)\|^2\right]$*
5. *$S_{\tau_t(\delta)-1}(\delta) \overset{a.s.}{\leq} {}^{\mathbb{E}[S_{t-1}(\delta)]}/_{\delta}$.*
6. *$\tau_{\tau_t(\delta)-1}(\delta) \overset{a.s.}{=} \tau_t(\delta) - 1$*
7. *For every $s < \tau_t(\delta)$, the following inequalities hold deterministically:*

$$\tilde{\eta}_s \geq \frac{\eta}{\sqrt{b_0^2 + 2\eta^2 L_0^2 + \sigma_0^2 + S_{\tau_t(\delta)-1}(\delta)}} \geq \frac{\eta}{\sqrt{b_0^2 + 2\eta^2 L_0^2 + \frac{(t-1)\sigma_0^2 + (1+\sigma_1^2+c_{\mathrm{L}})\mathbb{E}\left[\sum_{\ell<\tau_{t-1}(\delta)}\|\nabla F(\mathbf{w}_\ell)\|^2\right]}{\delta}}}$$

8. *$t \geq \mathbb{E}\left[\tau_t(\delta)\right] \geq t\left(1 - {}^{\delta(t-1)}/_{2}\right)$*

Before proving this result, let us briefly discuss an alternative construction to Definition 9 which is (perhaps) more natural and easier to define, but does not satisfy a property we rely on to prove Lemma 25:

Remark 24 *One might attempt to define the stopping times $\tau_t(\delta)$ from Definition 9 in the following simpler manner. First, denote $\widetilde{S}_t = \sum_{s=1}^{t-1} \|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}} \|\nabla F(\mathbf{w}_s)\|^2$. Then, let $\widetilde{\tau}_t(\delta) := \min\left\{t, \min\left\{s \geq 0 : \widetilde{S}_t > \mathbb{E}[\widetilde{S}_t]/\delta\right\}\right\}$. On a first impression, this stopping time might seem to capture the same properties as Definition 9. Unfortunately, this is not the case. To see this, let us examine the quantity $\widetilde{S}_{\widetilde{\tau}_t(\delta)-1}$. This stopping time guarantees the following:*

$$\begin{aligned}\widetilde{S}_{\widetilde{\tau}_t(\delta)-1} = \sum_{\ell=1}^{t-1} \widetilde{S}_\ell(\delta)\mathbb{1}\{\widetilde{\tau}_t(\delta)-1=\ell\} &\leq \sum_{\ell=1}^{t-1} \frac{\mathbb{E}\left[\widetilde{S}_\ell\right]}{\delta}\mathbb{1}\{\widetilde{\tau}_t(\delta)-1=\ell\} \\ &= \sum_{\ell=1}^{t-1} \frac{\sum_{s=1}^{\ell}\mathbb{E}\left[\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2\right]}{\delta}\mathbb{1}\{\widetilde{\tau}_t(\delta)-1=\ell\} \\ &= \frac{\sum_{s=1}^{\widetilde{\tau}_t(\delta)-1}\mathbb{E}\left[\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2\right]}{\delta}. \end{aligned} \tag{7}$$

Thus, we are only guaranteed deterministically that $\widetilde{S}_{\widetilde{\tau}_t(\delta)-1} \leq \sum_{\ell=1}^{t-1}\mathbb{E}\left[\|\boldsymbol{g}_\ell\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_\ell)\|^2\right]/\delta$ (indeed, this is the only inequality we know on any sample path where $\widetilde{\tau}_t(\delta) = t$). By contrast, by Item 5 of Lemma 23, we know that, deterministically:

$$S_{\tau_t(\delta)-1}(\delta) \leq \frac{\mathbb{E}\left[S_{t-1}(\delta)\right]}{\delta} = \frac{\mathbb{E}\left[\sum_{s=1}^{\tau_{t-1}(\delta)-1}\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2\right]}{\delta} \tag{8}$$

Notice that (8) is true no matter the realization of $\tau_t(\delta)$. Indeed, for any realization of $\tau_t(\delta)$, the bound on the right-hand side still involves a random index inside *the expectation. This is not the case with (7) (there, the random index is* outside *of the expectation). This difference is crucial, and this special property of $S_{\tau_t(\delta)-1}$ is actually what makes the proof of Lemma 25 possible.*

Proof (of Lemma 23) We prove the first claim via induction. The base case of $t = 1$ holds trivially, since $X_0(\delta) = 1$ deterministically by definition, which implies that $\tau_1(\delta) = 1$ and $S_1(\delta) = 0$, and thus $X_1(\delta) = 1$ (so are all measurable in the trivial σ-algebra). Assuming the claim holds for times $1, \ldots, t$, then we have that $\tau_{t+1}(\delta)$ is $\mathcal{F}_{t-1}$-measurable, since it depends only on $X_0(\delta), \ldots, X_t(\delta)$, each of which is $\mathcal{F}_{t-1}$-measurable by the induction hypothesis. Thus, since $S_{t+1}(\delta)$ depends only on $\tau_{t+1}(\delta)$ and $\left\{\|\boldsymbol{g}_s\|^2, \|\nabla F(\mathbf{w}_s)\|^2\right\}_{s=1}^{\tau_{t+1}(\delta)-1} \subseteq \left\{\|\boldsymbol{g}_s\|^2, \|\nabla F(\mathbf{w}_s)\|^2\right\}_{s=1}^{t}$, $S_{t+1}(\delta)$ is $\mathcal{F}_t$-measurable. Further, since $X_t(\delta)$ is $\mathcal{F}_{t-1} \subset \mathcal{F}_t$-measurable and $S_{t+1}(\delta)$ is $\mathcal{F}_t$-measurable, and by definition, $X_{t+1}(\delta) = X_t(\delta)\mathbb{1}\{S_{t+1}(\delta) \leq \mathbb{E}[S_{t+1}(\delta)]/\delta\}$, we conclude that $X_{t+1}(\delta)$ is $\mathcal{F}_t$-measurable. Thus, the claim holds by induction.

For the second claim, it suffices to consider $0 \leq s \leq t-2$ (since we just established that $\tau_t(\delta)$ is $\mathcal{F}_{t-2}$-measurable, and $\mathcal{F}_{t-2} \subset \mathcal{F}_{t'-2}$ for any $t' \geq t$). Now, for any such s, since $s < t$, we have that

$$\{s \geq \tau_t(\delta)\} = \cup_{\ell=0}^{s}\{X_\ell(\delta) = 0\} \in \mathcal{F}_{s-1},$$

since $X_\ell(\delta)$ is $\mathcal{F}_{s-1}$-measurable for every $\ell \leq s$. Thus, since $\mathcal{F}_{s-1}$ is a σ-algebra, and hence closed under complements, $\{s < \tau_t(\delta)\} = \{s \geq \tau_t(\delta)\}^c \in \mathcal{F}_{s-1}$.

For the third claim, the inequality $\tau_{t+1}(\delta) \geq \tau_t(\delta)$ follows immediately from the definition, since if $\tau_t(\delta) = s$ for some $s \in [t]$, then either $X_s = 0$, in which case $\tau_{t+1}(\delta) = s = \tau_t(\delta)$,

or $s = t$ and $X_t = 1$, in which case $\tau_t(\delta) = t$ and $\tau_{t+1}(\delta) = t + 1 > \tau_t(\delta)$. The inequality $S_{t+1}(\delta) \geq S_t(\delta)$ follows since $\tau_{t+1}(\delta) \geq \tau_t(\delta)$ and $S_t(\delta)$ is a sum of non-negative terms over the interval $[1, \tau_t(\delta))$, each of which is contained in the sum $S_{t+1}(\delta)$. The inequality $X_{t+1}(\delta) \leq X_t(\delta)$ follows immediately from the definition, since $\mathbb{1}\{S_{t+1}(\delta) \leq \mathbb{E}[S_{t+1}(\delta)]/\delta\} \in \{0, 1\}$.

For the fourth claim, we have that, by definition of $S_t(\delta)$ and the tower rule of expectation,

$$\begin{aligned}\mathbb{E}\left[S_t(\delta)\right] &= \mathbb{E}\left[\sum_{s=1}^{t-1}(\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2)\mathbb{1}\{s < \tau_t(\delta)\}\right] \\ &= \sum_{s=0}^{t-1}\mathbb{E}\left[\mathbb{E}\left[(\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2)\mathbb{1}\{s < \tau_t(\delta)\} \mid \mathcal{F}_{s-1}\right]\right].\end{aligned}$$

Now, since $\{s < \tau_t(\delta)\} \in \mathcal{F}_{s-1}$, and applying (1),

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[(\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2)\mathbb{1}\{s < \tau_t(\delta)\} \mid \mathcal{F}_{s-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{g}_s\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_s)\|^2 \mid \mathcal{F}_{s-1}\right]\mathbb{1}\{s < \tau_t(\delta)\}\right] \\ &\leq \mathbb{E}\left[(\sigma_0^2 + (1 + \sigma_1^2 + c_{\mathrm{L}})\|\nabla F(\mathbf{w}_s)\|^2)\mathbb{1}\{s < \tau_t(\delta)\}\right].\end{aligned}$$

Summing the above expression over $s \in [t-1]$, we conclude that

$$\mathbb{E}\left[S_t(\delta)\right] \leq \mathbb{E}\left[\sum_{s<\tau_t(\delta)} \sigma_0^2 + (1 + \sigma_1^2 + c_{\mathrm{L}})\|\nabla F(\mathbf{w}_s)\|^2\right],$$

establishing the third claim.

For the fifth claim, notice that, by definition of $\tau_t(\delta)$, if $\tau_t(\delta) = s$, then $X_{s-1}(\delta) = 1$, which implies that $S_{s-1}(\delta) \leq \mathbb{E}[S_{s-1}(\delta)]/\delta$ by construction. Therefore,

$$\begin{aligned}S_{\tau_t(\delta)-1}(\delta) = \sum_{s=1}^{t} S_{s-1}(\delta)\mathbb{1}\{\tau_t(\delta) = s\} &= \sum_{s=1}^{t} S_{s-1}(\delta)\mathbb{1}\{\tau_t(\delta) = s, S_{s-1}(\delta) \leq \mathbb{E}[S_{s-1}(\delta)]/\delta\} \\ &\leq \sum_{s=1}^{t}\frac{\mathbb{E}\left[S_{s-1}(\delta)\right]}{\delta}\mathbb{1}\{\tau_t(\delta) = s\} \leq \frac{\mathbb{E}\left[S_{t-1}(\delta)\right]}{\delta}.\end{aligned}$$

For the sixth claim, we note that

$$\begin{aligned}\tau_{\tau_t(\delta)-1}(\delta) &= \sum_{s=1}^{t}\tau_{s-1}(\delta)\mathbb{1}\{\tau_t(\delta) = s\} = \sum_{s=1}^{t}\tau_{s-1}(\delta)\mathbb{1}\{X_0(\delta) = \ldots = X_{s-1}(\delta) = 1, \tau_t(\delta) = s\} \\ &= \sum_{s=1}^{t}(s-1)\mathbb{1}\{\tau_t(\delta) = s\} = \sum_{s=0}^{t-1} s\mathbb{1}\{\tau_t(\delta) - 1 = s\} = \tau_t(\delta) - 1.\end{aligned}$$

For the seventh claim, assuming $s < \tau_t(\delta)$, we have that, by Lemma 18,

$$\begin{aligned}\tilde{\eta}_s &= \frac{\eta}{\sqrt{b_0^2 + \sigma_0^2 + \sum_{\ell=1}^{s-1}\|\boldsymbol{g}_\ell\|^2 + \|\nabla F(\mathbf{w}_s)\|^2}} \\ &\geq \frac{\eta}{\sqrt{b_0^2 + \sigma_0^2 + 2\eta^2 L_0^2 + \sum_{\ell<\tau_t(\delta)-1}\|\boldsymbol{g}_\ell\|^2 + c_{\mathrm{L}}\|\nabla F(\mathbf{w}_\ell)\|^2}}.\end{aligned}$$

Further, since $\tau_{\tau_t(\delta)-1}(\delta) = \tau_t(\delta) - 1$, and by definition of $S_t(\delta)$, we have that

$$\begin{aligned}\tilde{\eta}_s &\geq \frac{\eta}{\sqrt{b_0^2 + \sigma_0^2 + 2\eta^2 L_0^2 + \sum_{\ell < \tau_{\tau_t(\delta)-1}(\delta)} \|\boldsymbol{g}_\ell\|^2 + c_{\mathrm{L}} \|\nabla F(\mathbf{w}_\ell)\|^2}} \\ &= \frac{\eta}{\sqrt{b_0^2 + \sigma_0^2 + 2\eta^2 L_0^2 + S_{\tau_t(\delta)-1}(\delta)}}.\end{aligned}$$

Therefore, since $S_{\tau_t(\delta)-1}(\delta) \leq {}^{\mathbb{E}[S_{t-1}(\delta)]}/_{\delta}$ almost surely by Item 5, together with our upper-bound on $\mathbb{E}[S_t(\delta)]$ from Item 4, we conclude that

$$\tilde{\eta}_s \geq \frac{\eta}{\sqrt{b_0^2 + 2\eta^2 L_0^2 + \frac{(t-1)\sigma_0^2 + (1+\sigma_1^2+c_{\mathrm{L}})\mathbb{E}\left[\sum_{\ell < \tau_{t-1}(\delta)} \|\nabla F(\mathbf{w}_\ell)\|^2\right]}{\delta}}},$$

as claimed.

For the final claim, we note that $\tau_t(\delta) \leq t$ deterministically, by construction. Thus, we focus on the lower bound. Indeed, notice that, since $\tau_t(\delta) \in [t]$,

$$\begin{aligned}\tau_t(\delta) = \sum_{s=1}^{t} s\mathbb{1}\{\tau_t(\delta) = s\} &= \sum_{s=1}^{t} \mathbb{1}\{\tau_t(\delta) = s\} \sum_{\ell=0}^{s-1} \mathbb{1}\{\tau_t(\delta) > \ell\} \\ &= \sum_{\ell=0}^{t-1} \sum_{s=\ell+1}^{t} \mathbb{1}\{\tau_t(\delta) = s\}\mathbb{1}\{\tau_t(\delta) > \ell\} \\ &= \sum_{\ell=0}^{t-1} \mathbb{1}\{\tau_t(\delta) > \ell\} \sum_{s=\ell+1}^{t} \mathbb{1}\{\tau_t(\delta) = s\} = \sum_{\ell=0}^{t-1} \mathbb{1}\{\tau_t(\delta) > \ell\}\end{aligned}$$

Next, notice that $X_s(\delta) = 1$ iff $\tau_t(\delta) > s$, which implies that $X_s(\delta) = \mathbb{1}\{\tau_t(\delta) > s\}$. Additionally, recall that $X_0(\delta) = 1$, and $X_s(\delta) = \mathbb{1}\{\cap_{\ell=1}^{s} \{S_\ell(\delta) \leq {}^{\mathbb{E}[S_\ell(\delta)]}/_{\delta}\}\}$. Hence, we have that

$$\begin{aligned}\mathbb{E}[\tau_t(\delta)] = \sum_{s=0}^{t-1} \mathbb{E}[X_s(\delta)] = \sum_{s=0}^{t-1} \Pr[X_s(\delta) = 1] &= 1 + \sum_{s=1}^{t-1} 1 - \Pr[X_s(\delta) = 0] \\ &= 1 + \sum_{s=1}^{t-1} 1 - \Pr\left[\cup_{\ell=1}^{s} \{S_\ell(\delta) > {}^{\mathbb{E}[S_\ell(\delta)]}/_{\delta}\}\right].\end{aligned}$$

Therefore, by applying the union bound and Markov's inequality, we conclude that

$$\mathbb{E}[\tau_t(\delta)] \geq t - \sum_{s=1}^{t-1}\sum_{\ell=1}^{s} \Pr[S_\ell(\delta) > {}^{\mathbb{E}[S_\ell(\delta)]}/_{\delta}] \geq t - \sum_{s=1}^{t-1}\sum_{\ell=1}^{s} \delta = t - \delta\frac{t(t-1)}{2} = t\left(1 - \frac{\delta(t-1)}{2}\right),$$

which establishes the final claim. ■

B.3. The key consequence of the nice stopping time construction

The following result is the most crucial place where the properties of Definition 9 are utilized. It tells us that, as long as the sum of "bad" gradients is comparable to the sum of "good" ones, and as long as the descent inequality (Lemma 8) holds, then the sum of gradients scales (roughly) as $\mathcal{O}(b(T)^2/\delta + b(T)\sqrt{T/\delta})$. One can compare this result to that of Faw et al. (2022, Lemma 13), which obtained a similar bound in the simpler L_0-smooth setting. Their argument utilized a technique they termed "recursive improvement," which required recursively invoking gradually improving bounds in order to reach their desired conclusion after infinitely many calls. Moreover, their argument crucially relies on properties of L_0-smoothness in order to obtain worst-case upper bounds on the sum of gradients, which are no longer true in our setting. Through our construction of the stopping time $\tau_{T+1}(\delta)$, we are able to obtain a similar bound as in their setting, but with an (arguably) significantly simpler and more general proof which works even in the (L_0, L_1)-smooth setting.

Lemma 25 *Recall the stopping time $\tau_{T+1}(\delta)$ from Definition 9 and the set of "good" times before $\tau_{T+1}(\delta)$, $S_{\text{good}}(\tau_{T+1}(\delta))$ from Definition 21. Let $\widetilde{S}(\tau_{T+1}(\delta)) \subseteq S_{\text{good}}(\tau_{T+1}(\delta))$ be any (random) subset. Suppose that the following two conditions are satisfied: (i) for some $c_{\text{B1}}, c_{\text{B2}} \geq 0$ (possibly dependent on T):*

$$\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))^c} \|\nabla F(\mathbf{w}_t)\|^2\right] \leq c_{\text{B1}} + c_{\text{B2}}\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2\right] \tag{9}$$

and (ii) for some $b(T) \geq 0$,

$$\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right] \leq b(T). \tag{10}$$

Then, we obtain the inequality given below:

$$\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2\right] \leq \frac{2(1+c_{\text{B2}})(1+c_{\text{L}}+\sigma_1^2)b(T)^2}{\eta^2\delta} + \frac{2b(T)}{\eta}\sqrt{b_0^2 + 2\eta^2 L_0^2 + \frac{T\sigma_0^2 + (1+c_{\text{L}}+\sigma_1^2)c_{\text{B1}}}{\delta}}.$$

Proof Let $\widetilde{S}(\tau_{T+1}(\delta)) \subseteq S_{\text{good}}(\tau_{T+1}(\delta))$ be any (possibly random) subset. By (10),

$$b(T) \geq \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2\right].$$

Now, by Item 7 in Lemma 23, since $t < \tau_{T+1}(\delta)$ for any $t \in \widetilde{S}(\tau_{T+1}(\delta))$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] \\
&\geq \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \frac{\eta \left\|\nabla F(\mathbf{w}_t)\right\|^2}{\sqrt{b_0^2 + 2\eta L_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2+c_{\mathrm{L}})\mathbb{E}\left[\sum_{\ell<\tau_{T+1}(\delta)}\left\|\nabla F(\mathbf{w}_\ell)\right\|^2\right]}{\delta}}}\right] \\
&= \frac{\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \eta \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]}{\sqrt{b_0^2 + 2\eta L_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2+c_{\mathrm{L}})\mathbb{E}\left[\sum_{\ell<\tau_{T+1}(\delta)}\left\|\nabla F(\mathbf{w}_\ell)\right\|^2\right]}{\delta}}} \\
&= \frac{\eta E_{\mathrm{good}}}{\sqrt{b_0^2 + 2\eta L_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2+c_{\mathrm{L}})(E_{\mathrm{good}}+E_{\mathrm{bad}})}{\delta}}},
\end{aligned}$$

where $E_{\mathrm{good}} = \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]$ and $E_{\mathrm{bad}} = \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))^c} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right]$. Rearranging, we have the following inequality:

$$\eta E_{\mathrm{good}} \leq \sqrt{b_0^2 + 2\eta^2 L_0^2 + \frac{T\sigma_0^2 + (1+c_{\mathrm{L}}+\sigma_1^2)E_{\mathrm{bad}} + (1+c_{\mathrm{L}}+\sigma_1^2)E_{\mathrm{good}}}{\delta}}\, b(T).$$

Notice that this is a quadratic inequality in $\sqrt{E_{\mathrm{good}}}$. Assuming that

$$E_{\mathrm{bad}} \leq c_{\mathrm{B1}} + c_{\mathrm{B2}} E_{\mathrm{good}},$$

then we may solve this inequality to conclude that

$$\begin{aligned}
\sqrt{E_{\mathrm{good}}} &\leq \frac{\sqrt{(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)}\, b(T)}{2\eta\sqrt{\delta}} \\
&\quad + \frac{1}{2\eta}\sqrt{\frac{(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)^2}{\delta} + 4\eta\sqrt{b_0^2 + 2\eta^2 L_0^2 + \frac{T\sigma_0^2 + (1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}{\delta}}\, b(T)} \\
&\leq \frac{\sqrt{(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)}\, b(T)}{\eta\sqrt{\delta}} + \sqrt{\frac{b(T)}{\eta}}\sqrt[4]{b_0^2 + 2\eta^2 L_0^2 + \frac{T\sigma_0^2 + (1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}{\delta}},
\end{aligned}$$

from which we conclude

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \left\|\nabla F(\mathbf{w}_t)\right\|^2\right] = E_{\mathrm{good}} &\leq \frac{2(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)^2}{\eta^2\delta} \\
&\quad + \frac{2b(T)}{\eta}\sqrt{b_0^2 + 2\eta^2 L_0^2 + \frac{T\sigma_0^2 + (1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}{\delta}}.
\end{aligned}$$

■

B.4. Convergence for (L_0, L_1)-smooth functions

Here, we provide our main theorem for (L_0, L_1)-smooth functions. We emphasize that, unlike in the statement of Theorem 3 from the main body, this theorem does not (directly) require $\sigma_1 < 1$. Instead, it requires that $\sum_{t \in \widetilde{S}(\tau_{T+1}(\delta))^c} \|\nabla F(\mathbf{w}_t)\|^2$, $\mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right]$, and $\mathrm{comp}(\tau_{T+1})$ can each be upper-bounded by sufficiently-small quantities. While these quantities can each be (trivially) upper-bounded when $\sigma_1 < 1$, this is not a necessary condition. Indeed, we prove in Corollary 32 convergence for a subset of (L_0, L_1)-smooth functions without a restriction on σ_1 using this theorem as well.

Theorem 26 (Formal statement of Theorem 3) *Fix any $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$. Consider* (AG-Norm) *with any parameters $\eta \leq {}^{2\varepsilon'}/_{L_1(4+\sigma_1^2)}$ and $b_0^2 > 0$, running for $T \geq 1$ time steps on an objective function satisfying Assumption 2, and given access to a stochastic gradient oracle satisfying Assumptions 3 and 4. Let, for any $\delta' \in (0,1)$, $\tau_{T+1} := \tau_{T+1}({}^{\delta'}/_{4T})$ be the stopping time from Definition 9. Let $S_{\mathrm{good}}(\tau_{T+1})$ by the set of "good times" from Definition 21, let $\widetilde{S}(\tau_{T+1}) \subseteq S_{\mathrm{good}}(\tau_{T+1})$, and denote $S^{\mathrm{comp}}(\tau_{T+1}) := S_{\mathrm{good}}(\tau_{T+1}) \setminus \widetilde{S}(\tau_{T+1})$ to be the compensating "good" times for the bad times $S_{\mathrm{good}}(\tau_{T+1})^c$. Suppose there is a (possibly random) $\mathrm{B}_1 \geq 0$ and constant $c_{\mathrm{B2}} \geq 0$ such that $\mathbb{E}[\mathrm{B}_1] \leq c_{\mathrm{B1}} < \infty$ and which (deterministically) satisfy:*

$$\sum_{t \in \widetilde{S}(\tau_{T+1})^c} \|\nabla F(\mathbf{w}_t)\|^2 \leq \mathrm{B}_1 + c_{\mathrm{B2}} \sum_{t \in \widetilde{S}(\tau_{T+1})} \|\nabla F(\mathbf{w}_t)\|^2 .$$

then for any $T \geq 1$ and $\delta' \in (0,1)$, with probability at least $1 - \delta' - {}^{2\mathbb{E}[|\widetilde{S}(\tau_{T+1})^c|]}/_{T}$, (AG-Norm) *satisfies:*

$$\begin{aligned}
&\min_{t \in [T]} \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \frac{32(1+c_{\mathrm{B2}})b(T)^2}{\eta^2(\delta')^2 T} + \frac{16b(T)}{\eta(\delta')^2 T}\sqrt{b_0^2 + \sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + \frac{4b(T)}{\eta}\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{b_0^2 + 2\eta^2 L_0^2}} \\
&\quad + \frac{32 b(T)^{3/2}}{\eta^{3/2}(\delta')^{2.25} T^{3/4}}\sqrt{2\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{(1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}} \\
&\quad + \frac{16 b(T)}{\eta(\delta')^2\sqrt{T}}\sqrt{2\sigma_0^2 + \frac{8\sigma_1^2(1+c_{\mathrm{B2}})b(T)}{\eta\sqrt{\delta'}}\left(\frac{2(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)}{\eta\sqrt{\delta'}} + \sigma_0\right)},
\end{aligned}$$

where $c_{\mathrm{L}} = 2(1+\eta L_1)^2$,

$$b(T) := \frac{1}{\varepsilon'''}\left(F(\mathbf{w}_1) - F^* + 2\widetilde{c}_0 \log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0 \mathbb{E}\left[\tau_{T+1} - 1\right]}{\eta\varepsilon'' b_0}\right) + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)} + \mathrm{comp}(\tau_{T+1})\right),$$

and

$$\mathrm{comp}(\tau_{T+1}) = \mathbb{E}\left[\sum_{t \in S_{\mathrm{good}}(\tau_{T+1})^c} (\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t' \in S^{\mathrm{comp}}(\tau_{T+1})} \varepsilon''' \tilde{\eta}_{t'} \|\nabla F(\mathbf{w}_{t'})\|^2\right],$$

and $\widetilde{c}_0 = \frac{\eta\sigma_0}{2\varepsilon} + \eta^2 \frac{L_0 + \sigma_0 L_1}{2}$.

In particular, whenever $\sigma_1 \leq (1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))$*, then we have that* $S_{\text{good}}(\tau_{T+1}) = [\tau_{T+1}-1]$*, so we can take* $\widetilde{S}(\tau_{T+1}) = S_{\text{good}}(\tau_{T+1})$ *so that* $c_{\text{B1}} = 0 = c_{\text{B2}}$ *and* $\text{comp}(\tau_{T+1}) \leq 0$*, so, with probability at least* $1-\delta'$*, the following inequality holds:*

$$\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 \leq \frac{32b(T)^2}{\eta^2(\delta')^2T} + \frac{16b(T)}{\eta(\delta')^2T}\sqrt{b_0^2+\sigma_0^2+\frac{4b(T)}{\eta}\sigma_1^2\sqrt{b_0^2+2\eta^2L_0^2}} + \frac{16b(T)}{\eta(\delta')^2\sqrt{T}}\sqrt{2\sigma_0^2+\frac{8\sigma_1^2b(T)}{\eta\sqrt{\delta'}}\left(\frac{2(1+c_{\text{L}}+\sigma_1^2)b(T)}{\eta\sqrt{\delta'}}+\sigma_0\right)},$$

Proof

Step 1: Rewrite Lemma 8 in terms of a single, worst-case step-size Let us assume that $\eta \leq {}^{2\varepsilon'}/_{L_1(4+\sigma_1^2)}$. Denote:

$$\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}} = \frac{\eta}{\sqrt{b_0^2+\sigma_0^2+2\sum_{s<\tau_{T+1}(\delta)}\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2+\|\nabla F(\mathbf{w}_s)\|^2}}.$$

Let $\widetilde{S}(\tau_{T+1}(\delta)) \subseteq S_{\text{good}}(\tau_{T+1}(\delta))$. Then, taking $b(T)$ as in the theorem statement above,

$$\mathbb{E}\left[\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2\right] \leq b(T), \tag{11}$$

since, by Lemma 8, and using the fact that $t < \tau_{T+1}(\delta)$ for every $t \in \widetilde{S}(\tau_{T+1}(\delta))$,

$$\begin{aligned}
b(T) &\geq \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2\right] \\
&= \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\frac{\eta\|\nabla F(\mathbf{w}_t)\|^2}{\sqrt{b_0^2+\sigma_0^2+\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2+\|\nabla F(\mathbf{w}_t)\|^2}}\right] \\
&\geq \mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\frac{\eta\|\nabla F(\mathbf{w}_t)\|^2}{\sqrt{b_0^2+\sigma_0^2+\sum_{s=1}^{t-1}(2\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2+2\|\nabla F(\mathbf{w}_s)\|^2)+\|\nabla F(\mathbf{w}_t)\|^2}}\right] \\
&\geq \mathbb{E}\left[\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2\right].
\end{aligned}$$

Step 2: Upper bound the inverse of this worst-case step-size in expectation Next, notice that, denoting $E_{\text{good}} = \mathbb{E}\left[\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2\right]$,

$$\mathbb{E}\left[{}^{\eta}/_{\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}}\right] \leq \sqrt{b_0^2+(2T+1)\sigma_0^2+2(1+\sigma_1^2)c_{\text{B1}}+2E_{\text{good}}} + \sqrt{2(1+c_{\text{B2}})}\mathbb{E}\left[\sqrt{\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2}\right], \tag{12}$$

since, by the assumption that $\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))^c}\|\nabla F(\mathbf{w}_s)\|^2 \le \mathrm{B}_1 + c_{\mathrm{B}2}\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2$ and $\mathbb{E}[\mathrm{B}_1]\le c_{\mathrm{B}1}$,

$$\begin{aligned}
\mathbb{E}\left[\eta/\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}\right] &= \mathbb{E}\left[\sqrt{b_0^2+\sigma_0^2+2\sum_{s<\tau_{T+1}(\delta)}\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2+\|\nabla F(\mathbf{w}_s)\|^2}\right]\\
&\le \mathbb{E}\left[\sqrt{b_0^2+\sigma_0^2+2\mathrm{B}_1+2\sum_{s<\tau_{T+1}(\delta)}\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2+2(1+c_{\mathrm{B}2})\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2}\right]\\
&\le \mathbb{E}\left[\sqrt{b_0^2+\sigma_0^2+2\mathrm{B}_1+2\sum_{s<\tau_{T+1}(\delta)}\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2}\right]\\
&\quad+\sqrt{2(1+c_{\mathrm{B}2})}\mathbb{E}\left[\sqrt{\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2}\right]\\
&\le \sqrt{b_0^2+\sigma_0^2+2c_{\mathrm{B}1}+2\mathbb{E}\left[\sum_{s<\tau_{T+1}(\delta)}\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2\right]}\\
&\quad+\sqrt{2(1+c_{\mathrm{B}2})}\mathbb{E}\left[\sqrt{\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2}\right]\\
&\le \sqrt{b_0^2+(2T+1)\sigma_0^2+2(1+\sigma_1^2)c_{\mathrm{B}1}+2(1+c_{\mathrm{B}2})\mathbb{E}\left[\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2\right]}\\
&\quad+\sqrt{2(1+c_{\mathrm{B}2})}\mathbb{E}\left[\sqrt{\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2}\right],
\end{aligned}$$

where, in the last step, we used the fact that, since $\{s<\tau_{T+1}(\delta)\}\in\mathcal{F}_{s-1}$ by Item 2 of Lemma 23, we may apply Assumption 4 to obtain

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2\mathbb{1}\{s<\tau_{T+1}(\delta)\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{g}_s-\nabla F(\mathbf{w}_s)\|^2\mid\mathcal{F}_{t-1}\right]\mathbb{1}\{s<\tau_{T+1}(\delta)\}\right]\\
&\le \mathbb{E}\left[\sigma_0^2+\sigma_1^2\|\nabla F(\mathbf{w}_s)\|^2\mathbb{1}\{s<\tau_{T+1}(\delta)\}\right].
\end{aligned}$$

Step 3: Use Hölder's inequality and the above bounds to obtain a quadratic inequality At this point, we can combine the bounds from (11) and (12) to obtain a quadratic inequality in $\mathbb{E}\left[\sqrt{\sum_{s\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_s)\|^2}\right]$. Indeed, using Hölder's inequality $\mathbb{E}[X^2]\ge\mathbb{E}[XY]^2/\mathbb{E}[Y^2]$ with $X=\sqrt{\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2}$ and $Y=\sqrt{1/\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}}$, we have that:

$$\frac{\mathbb{E}\left[\sqrt{\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2}\right]^2}{\mathbb{E}\left[1/\underline{\tilde{\eta}_{\tau_{T+1}(\delta)}}\right]}\le b(T),$$

which, after rearranging, and writing $Z = \sqrt{\sum_{s \in \widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_s)\|^2}$, implies that:

$$\eta \mathbb{E}\left[Z\right]^2 \leq b(T)\left(\sqrt{b_0^2 + (2T+1)\sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + 2\sigma_1^2 E_{\mathrm{good}}} + \sqrt{2(1+c_{\mathrm{B2}})}\mathbb{E}\left[Z\right]\right). \tag{13}$$

We solve this quadratic inequality in $\mathbb{E}\left[Z\right]$ to conclude that:

$$\begin{aligned}
\mathbb{E}\left[Z\right] &\leq \frac{\sqrt{2(1+c_{\mathrm{B2}})}b(T)}{2\eta} \\
&\quad + \frac{1}{2\eta}\sqrt{2(1+c_{\mathrm{B2}})b(T)^2 + 4\eta b(T)\sqrt{b_0^2 + (2T+1)\sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + 2\sigma_1^2(1+c_{\mathrm{B2}})E_{\mathrm{good}}}} \\
&\leq \frac{\sqrt{2(1+c_{\mathrm{B2}})}b(T)}{\eta} + \sqrt{\frac{b(T)}{\eta}}\sqrt[4]{b_0^2 + (2T+1)\sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + 2\sigma_1^2(1+c_{\mathrm{B2}})E_{\mathrm{good}}}.
\end{aligned}$$

Thus, applying the bound on E_{good} from Lemma 25, we obtain:

$$\begin{aligned}
\mathbb{E}\left[Z\right]^2 &\leq \frac{4(1+c_{\mathrm{B2}})b(T)^2}{\eta^2} + \frac{2b(T)}{\eta}\sqrt{b_0^2 + \sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + \frac{4b(T)}{\eta}\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{b_0^2 + 2\eta^2 L_0^2}} \\
&\quad + \frac{2b(T)}{\eta}\sqrt{2T\sigma_0^2 + \frac{4\sigma_1^2(1+c_{\mathrm{B2}})b(T)}{\eta\sqrt{\delta}}\left(\frac{(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)}{\eta\sqrt{\delta}} + \sqrt{T\sigma_0^2 + (1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}\right)},
\end{aligned}$$

where $\delta \in (0,1)$ is a parameter of our choosing. In particular, choosing (with foresight) $\delta = \delta'/4T$ for any $\delta' \in (0,1)$, the above can be rewritten as:

$$\begin{aligned}
\mathbb{E}\left[Z\right]^2 &\leq \frac{4(1+c_{\mathrm{B2}})b(T)^2}{\eta^2} + \frac{2b(T)}{\eta}\sqrt{b_0^2 + \sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + \frac{4b(T)}{\eta}\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{b_0^2 + 2\eta^2 L_0^2}} \\
&\quad + \frac{4b(T)^{3/2}\sqrt[4]{T}}{\eta^{3/2}(\delta')^{1/4}}\sqrt{2\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{(1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}} \\
&\quad + \frac{2b(T)\sqrt{T}}{\eta}\sqrt{2\sigma_0^2 + \frac{8\sigma_1^2(1+c_{\mathrm{B2}})b(T)}{\eta\sqrt{\delta'}}\left(\frac{2(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)}{\eta\sqrt{\delta'}} + \sigma_0\right)} \\
&:= C_T^2
\end{aligned}$$

Step 4: Use the conclusion of the quadratic inequality to conclude with the claimed convergence guarantee To obtain a convergence rate, we begin by noting, for any $\delta' \in (0,1)$, we can decompose

$$\begin{aligned}
\Pr\left[\min_{t\in[T]}\|\nabla F(\mathbf{w}_t)\|^2 > \frac{8C_T^2}{(\delta')^2 T}\right] &= \Pr\left[\min_{t\in[T]}\|\nabla F(\mathbf{w}_t)\|^2 > \frac{8C_T^2}{(\delta')^2 T}, |\widetilde{S}(\tau_{T+1}(\delta))| \leq T/2\right] \\
&\quad + \Pr\left[\min_{t\in[T]}\|\nabla F(\mathbf{w}_t)\|^2 > \frac{8C_T^2}{(\delta')^2 T}, |\widetilde{S}(\tau_{T+1}(\delta))| > T/2\right]
\end{aligned} \tag{14}$$

The first term in (14) is easy to bound via Markov's inequality, since, choosing $\delta = \frac{\delta'}{4T} \leq \frac{\delta'}{2(T+1)}$ (since $T \geq 1$),

$$\begin{aligned}\Pr\left[\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 > \frac{8C_T^2}{(\delta')^2 T}, |\widetilde{S}(\tau_{T+1}(\delta))| \leq T/2\right] &\leq \Pr\left[|\widetilde{S}(\tau_{T+1}(\delta))| \leq T/2\right] \\ &\leq \frac{2\mathbb{E}\left[|[T] \setminus \widetilde{S}(\tau_{T+1}(\delta))|\right]}{T} \\ &\leq \frac{2}{T}\left(\frac{\delta T(T+1)}{2} + \mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right]\right) \\ &\leq \frac{\delta'}{2} + \frac{2\mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right]}{T}, \end{aligned} \tag{15}$$

where we bounded the above expectation using the fact that, by Item 8 of Lemma 23,

$$\begin{aligned}\mathbb{E}\left[|[T] \setminus \widetilde{S}(\tau_{T+1}(\delta))|\right] = \mathbb{E}\left[|[\tau_{T+1}(\delta), T] \cup \widetilde{S}(\tau_{T+1}(\delta))^c|\right] &\leq \mathbb{E}\left[T - \tau_{T+1}(\delta) + 1\right] + \mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right] \\ &\leq \frac{\delta T(T+1)}{2} + \mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right].\end{aligned}$$

To bound the second term in (14), we note that, whenever $|\widetilde{S}(\tau_{T+1}(\delta))| > T/2$, then

$$\begin{aligned}\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 \leq \min_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2 &\leq \frac{1}{|\widetilde{S}(\tau_{T+1}(\delta))|} \sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2 \\ &\leq \frac{2}{T} \underbrace{\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2}_{Z^2}\end{aligned}$$

Hence, we have by Markov's inequality and the above bound,

$$\Pr\left[\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 > \frac{8C_T^2}{(\delta')^2 T}, |\widetilde{S}(\tau_{T+1}(\delta))| > T/2\right] \leq \Pr\left[Z > \frac{2C_T}{\delta'}\right] \leq \delta' \frac{\mathbb{E}[Z]}{2C_T} \leq \frac{\delta'}{2}. \tag{16}$$

Therefore, combining (15) and (16) with (14), we conclude that, with probability at least $1 - \delta' - 2\mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right]/T$,

$$\begin{aligned}&\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 \\ &\leq \frac{8C_T^2}{(\delta')^2 T} \\ &\leq \frac{32(1+c_{\mathrm{B2}})b(T)^2}{\eta^2(\delta')^2 T} + \frac{16b(T)}{\eta(\delta')^2 T}\sqrt{b_0^2 + \sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + \frac{4b(T)}{\eta}\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{b_0^2 + 2\eta^2 L_0^2}} \\ &\quad + \frac{32b(T)^{3/2}}{\eta^{3/2}(\delta')^{2.25}T^{3/4}}\sqrt{2\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{(1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}} \\ &\quad + \frac{16b(T)}{\eta(\delta')^2\sqrt{T}}\sqrt{2\sigma_0^2 + \frac{8\sigma_1^2(1+c_{\mathrm{B2}})b(T)}{\eta\sqrt{\delta'}}\left(\frac{2(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)}{\eta\sqrt{\delta'}} + \sigma_0\right)},\end{aligned}$$

as claimed ■

B.5. A deferred proof for establishing Lemma 6

Here, we give a bound which is used in proving Lemma 6. We remark that this inequality is an extension of a similar one from (Faw et al., 2022) (in the L_0-smooth setting) to the more general (L_0, L_1)-smooth setting. We additionally note that this bound has a better dependence on $\sigma_1 \text{bias}_t$ than the analogous one in theirs.

Lemma 27 *Fix any $\varepsilon \in (0,1)$. Suppose that $\eta \leq 1/L_1$. Then, for any time t, the iterates of* (AG-Norm) *satisfy*

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] \leq -\tilde{\eta}_t \left(1 - \varepsilon - \sigma_1 \text{bias}_t\right) \|\nabla F(\mathbf{w}_t)\|^2 + c_0 \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] + \frac{L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right],$$

where

$$c_0 = \frac{\eta \sigma_0}{2\varepsilon} + \frac{\eta^2 L_0}{2} \quad \textit{and} \quad \text{bias}_t = \sqrt{\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right]}.$$

Proof The proof proceeds using similar arguments as in (Faw et al., 2022, Lemma 5). By Lemma 17 and the definition of (AG-Norm), we know that

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] &\leq -\mathbb{E}\left[\eta_t \langle \nabla F(\mathbf{w}_t), \boldsymbol{g}_t \rangle \mid \mathcal{F}_{t-1}\right] + \frac{L_0 + L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
&\leq -\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 - \mathbb{E}\left[(\eta_t - \tilde{\eta}_t) \langle \nabla F(\mathbf{w}_t), \boldsymbol{g}_t \rangle \mid \mathcal{F}_{t-1}\right] \\
&\quad + \frac{L_0 + L_1 \|\nabla F(\mathbf{w}_t)\|}{2} \mathbb{E}\left[\eta_t^2 \|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

We begin by bounding the inner product term above as:

$$-(\eta_t - \tilde{\eta}_t) \langle \nabla F(\mathbf{w}_t), \boldsymbol{g}_t \rangle \leq |\eta_t - \tilde{\eta}_t| \|\nabla F(\mathbf{w}_t)\| \|\boldsymbol{g}_t\|.$$

To bound this quantity, we begin by rewriting $\eta_t - \tilde{\eta}_t$. Denoting $\widetilde{b}_t^2 := b_t^2 + \|\widetilde{\nabla}_t\|^2$, we have that

$$|\eta_t - \tilde{\eta}_t| = \eta \left| \frac{1}{\sqrt{b_{t-1}^2 + \|\boldsymbol{g}_t\|^2}} - \frac{1}{\sqrt{b_{t-1}^2 + \|\widetilde{\nabla}_t\|^2}} \right| = \eta \frac{\left| \|\widetilde{\nabla}_t\|^2 - \|\boldsymbol{g}_t\|^2 \right|}{\widetilde{b}_t b_t (\widetilde{b}_t + \widetilde{b}_t)} = \eta \frac{\left| \|\widetilde{\nabla}_t\| - \|\boldsymbol{g}_t\| \right| \left(\|\widetilde{\nabla}_t\| + \|\boldsymbol{g}_t\| \right)}{\widetilde{b}_t b_t (\widetilde{b}_t + \widetilde{b}_t)}.$$

Combining the above arguments, and applying Hölder's inequality, we have that

$$\begin{aligned}
&- \mathbb{E}\left[(\eta_t - \tilde{\eta}_t) \langle \nabla F(\mathbf{w}_t), \boldsymbol{g}_t \rangle \mid \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[|\eta_t - \tilde{\eta}_t| \|\nabla F(\mathbf{w}_t)\| \|\boldsymbol{g}_t\| \mid \mathcal{F}_{t-1}\right] \\
&= \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\| \, \mathbb{E}\left[\frac{\|\boldsymbol{g}_t\| \left(\|\boldsymbol{g}_t\| + \|\widetilde{\nabla}_t\| \right)}{b_t (\widetilde{b}_t + b_t)} \left| \|\widetilde{\nabla}_t\| - \|\boldsymbol{g}_t\| \right| \mid \mathcal{F}_{t-1} \right] \\
&\leq \tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\| \sqrt{\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2 \left(\|\boldsymbol{g}_t\| + \|\widetilde{\nabla}_t\| \right)^2}{b_t^2 (\widetilde{b}_t + b_t)^2} \mid \mathcal{F}_{t-1} \right]} \sqrt{\mathbb{E}\left[\left| \|\widetilde{\nabla}_t\| - \|\boldsymbol{g}_t\| \right|^2 \mid \mathcal{F}_{t-1} \right]}.
\end{aligned}$$

By (1), $\mathbb{E}\left[\|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] \le \sigma_0^2 + (1+\sigma_1^2)\|\nabla F(\mathbf{w}_t)\|^2$, and by Assumption 3 and Jensen's inequality, $\mathbb{E}\left[\|\boldsymbol{g}_t\| \mid \mathcal{F}_{t-1}\right] \ge \|\mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right]\| = \|\nabla F(\mathbf{w}_t)\|$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\left|\|\widetilde{\nabla}_t\| - \|\boldsymbol{g}_t\|\right|^2 \mid \mathcal{F}_{t-1}\right] &= \|\widetilde{\nabla}_t\|^2 + \mathbb{E}\left[\|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right] - 2\|\widetilde{\nabla}_t\|\mathbb{E}\left[\|\boldsymbol{g}_t\| \mid \mathcal{F}_{t-1}\right] \\ &\le \|\widetilde{\nabla}_t\|^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla F(\mathbf{w}_t)\|^2 - 2\|\widetilde{\nabla}_t\|\,\|\nabla F(\mathbf{w}_t)\| \\ &\le 2\sigma_0^2 + \sigma_1^2\|\nabla F(\mathbf{w}_t)\|^2,\end{aligned}$$

where the last step comes from $\|\widetilde{\nabla}_t\| \ge \|\nabla F(\mathbf{w}_t)\|$. Collecting our bounds so far yields:

$$\begin{aligned}&-\mathbb{E}\left[(\eta_t - \tilde{\eta}_t)\langle \nabla F(\mathbf{w}_t), \boldsymbol{g}_t\rangle \mid \mathcal{F}_{t-1}\right] \\ &\le \tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|\sqrt{\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2(\|\widetilde{\nabla}_t\| + \|\boldsymbol{g}_t\|)^2}{b_t^2(\widetilde{b}_t + b_t)^2} \mid \mathcal{F}_{t-1}\right]}\sqrt{2\sigma_0^2 + \sigma_1^2\|\nabla F(\mathbf{w}_t)\|^2}\end{aligned}$$

Focusing on the term depending on σ_0, we have that for any $\varepsilon > 0$,

$$\begin{aligned}&\sqrt{2}\sigma_0\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|\sqrt{\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2\left(\|\widetilde{\nabla}_t\| + \|\boldsymbol{g}_t\|\right)^2}{b_t^2(\widetilde{b}_t + b_t)^2} \mid \mathcal{F}_{t-1}\right]} \\ &\le \varepsilon\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2 + \frac{\sigma_0^2\tilde{\eta}_t}{2\varepsilon}\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2\left(\|\widetilde{\nabla}_t\| + \|\boldsymbol{g}_t\|\right)^2}{b_t^2(\widetilde{b}_t + b_t)^2} \mid \mathcal{F}_{t-1}\right] \\ &\le \varepsilon\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2 + \frac{\sigma_0\eta}{2\varepsilon}\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right].\end{aligned}$$

Thus, denoting $\mathrm{bias}_t = \sqrt{\mathbb{E}\left[\|\boldsymbol{g}_t\|^2/b_t^2 \mid \mathcal{F}_{t-1}\right]}$ and $c_0 = \eta\sigma_0/2\varepsilon + \eta^2 L_0/2$, we have that

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \mid \mathcal{F}_{t-1}\right] \le -\tilde{\eta}_t(1 - \varepsilon - \sigma_1\mathrm{bias}_t)\|\nabla F(\mathbf{w}_t)\|^2 + c_0\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t-1}\right] \\ + \frac{L_1\|\nabla F(\mathbf{w}_t)\|}{2}\mathbb{E}\left[\eta_t^2\|\boldsymbol{g}_t\|^2 \mid \mathcal{F}_{t-1}\right],\end{aligned}$$

as claimed by the lemma. ■

Appendix C. Proofs for Polynomially-bounded functions for general σ_1

In this section, we show that Theorem 26 can be used to establish a $\widetilde{\mathcal{O}}(1/\sqrt{T})$ convergence rate without the restriction of $\sigma_1 < 1$. The key is to restrict our attention to (L_0, L_1)-smooth functions which satisfy the following additional property:

C.1. The key definition and its properties

Definition 4 *A function $F(\cdot)$ is* k-polynomially bounded *for $k \geq 2$ if $\forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, then there are constants $c_k \geq 1$ and $c_k', L_0 > 0$ such that:*

$$\|\nabla F(\mathbf{w})\| - c_k \left\|\nabla F(\mathbf{w}')\right\| \leq \max\left\{c_k' \left\|\mathbf{w} - \mathbf{w}'\right\|^{k-1}, L_0 \left\|\mathbf{w} - \mathbf{w}'\right\|\right\}.$$

The following result provides a characterization of these functions relative to L_0-smooth functions and (L_0, L_1)-smooth functions. In particular, it tells us that Definition 4 is a richer function class than (L_0, L_1)-smooth functions. However, not all (L_0, L_1)-smooth functions satisfy Definition 4.

Proposition 28 *We have the following:*

1. *Every L_0-smooth function satisfies Definition 4 with $k = 2$, $c_k = 1$, and $c_k' = L_0$.*
2. *Every (L_0, L_1)-smooth function satisfies Definition 4* locally *(i.e., when $\|\mathbf{w} - \mathbf{w}'\| \leq 1/L_1$) with $k = 2, c_k = 2$ and $c_k' = L_0$.*
3. *There is a $(0, L_1)$-smooth function which does not satisfy Definition 4 for any fixed k, c_k, c_k'.*
4. *For any $k \geq 2$, $F(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^*\|^k$ satisfies Definition 4 with $k = k$, $c_k = 2^{k-2}$, and $c_k' = k2^{k-2}$. Additionally, for any $L_1 > 0$, $F(\mathbf{w})$ is $(2k(k-1)/L_1^{k-2}, (e-1)(k-1)L_1)$-smooth. However, this $F(\mathbf{w})$ is not L_0-smooth when $k > 2$.*

In particular, this implies that:

$\{L_0$*-smooth functions*$\} \subsetneq \{(L_0, L_1)$*-smooth functions satisfying Definition 4*$\} \subsetneq \{(L_0, L_1)$*-smooth functions*$\}$

Proof The first claim follows by noting that L_0-smooth functions satisfy, for every $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$,

$$\|\nabla F(\mathbf{w})\| - \left\|\nabla F(\mathbf{w}')\right\| \leq \left|\|\nabla F(\mathbf{w})\| - \left\|\nabla F(\mathbf{w}')\right\|\right| \leq \left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\right\| \leq L_0 \left\|\mathbf{w} - \mathbf{w}'\right\|.$$

The second follows since, for any (L_0, L_1)-smooth function, for every $\|\mathbf{w} - \mathbf{w}'\| \leq L_1$,

$$\begin{aligned}\|\nabla F(\mathbf{w})\| - \left\|\nabla F(\mathbf{w}')\right\| \leq \left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\right\| &\leq \left(L_0 + L_1 \left\|\nabla F(\mathbf{w}')\right\|\right) \left\|\mathbf{w} - \mathbf{w}'\right\| \\ &\leq L_0 \left\|\mathbf{w} - \mathbf{w}'\right\| + \left\|\nabla F(\mathbf{w}')\right\|.\end{aligned}$$

For the third claim, consider the function $F(w) = \exp(L_1 w)$. Since $F''(w) = L_1^2 \exp(L_1 w) = L_1 F'(w)$. Suppose there were some k, c_k, c_k' such that Definition 4 is satisfied. Then, it must be the case that, for any $x > 0$:

$$\begin{aligned}1 \geq \lim_{\alpha \to \infty} \frac{\exp(L_1 \alpha x) - c_k \exp(L_1 0)}{c_k'(\alpha x)^{k-1}} &= \lim_{\alpha \to \infty} \frac{L_1 x \exp(L_1 \alpha x)}{c_k'(k-1)x^{k-1}\alpha^{k-2}} \\ &= \frac{L_1}{c_k'(k-1)x^{k-2}} \lim_{\alpha \to \infty} \frac{\exp(L_1 \alpha x)}{\alpha^{k-2}},\end{aligned}$$

where the inequality follows from the definition of Definition 4, the first equality by L'Hôpital's rule, and the second by rewriting the previous expression. Repeating this argument $k-1$ times, this implies that

$$1 \geq \lim_{\alpha \to \infty} \frac{\exp(L_1 \alpha x) - c_k \exp(L_1 0)}{c_k'(\alpha x)^{k-1}} = \frac{L_1^{k-1}}{c_k'(k-1)!} \lim_{\alpha \to \infty} \exp(L_1 \alpha x) = \infty,$$

a contradiction. Hence, $\exp(L_1 x)$ cannot satisfy Definition 4.

For the final claim, we see that $F(\mathbf{w})$ satisfies Definition 4 with $c_k = 2^{k-2}$ and $c_k' = k2^{k-2}$ since, by Jensen's inequality,

$$\begin{aligned}\|\nabla F(\mathbf{w})\| = k\|\mathbf{w}-\mathbf{w}^*\|^{k-1} &= k2^{k-1}\left\|\frac{1}{2}(\mathbf{w}-\mathbf{w}') + \frac{1}{2}(\mathbf{w}'-\mathbf{w}^*)\right\|^{k-1} \\ &\le k2^{k-2}(\|\mathbf{w}-\mathbf{w}'\|^{k-1} + \|\mathbf{w}'-\mathbf{w}^*\|^{k-1}) \\ &\le 2^{k-2}(k\|\mathbf{w}-\mathbf{w}'\|^{k-1} + \|\nabla F(\mathbf{w}')\|).\end{aligned}$$

Further, $F(\mathbf{w})$ is also $(2k(k-1), (e-1)(k-1))$-smooth, since simple calculations yield that

$$\nabla^2 F(\mathbf{w}) = k(k-2)\|\mathbf{w}-\mathbf{w}^*\|^{k-4}(\mathbf{w}-\mathbf{w}^*)(\mathbf{w}-\mathbf{w}^*)^\top + k\|\mathbf{w}-\mathbf{w}^*\|^{k-2} I.$$

In particular, this implies that $\mathbf{w}-\mathbf{w}^*$ is an eigenvector with largest eigenvalue, so, for any $L_1 > 0$,

$$\begin{aligned}\|\nabla^2 F(\mathbf{w})\| = k(k-1)\|\mathbf{w}-\mathbf{w}^*\|^{k-2} &\le k(k-1)\max\left\{L_1\|\mathbf{w}-\mathbf{w}^*\|^{k-1}, \frac{1}{L_1^{k-2}}\right\} \\ &\le \frac{k(k-1)}{L_1^{k-2}} + (k-1)L_1\|\nabla F(\mathbf{w})\|.\end{aligned}$$

Therefore, by (Zhang et al., 2020a, Corollary A.4), for any $\|\mathbf{w}-\mathbf{w}'\| \le 1/(k-1)L_1$,

$$\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \le \left(\frac{2k(k-1)}{L_1^{k-2}} + (e-1)(k-1)L_1\|\nabla F(\mathbf{w}')\|\right)\|\mathbf{w}-\mathbf{w}'\|.$$

Hence, F is $(2k(k-1)/L_1^{k-2}, (e-1)(k-1)L_1)$-smooth, as claimed. It is clear that this F is not L_0-smooth for any L_0 when $k > 2$, since for any $\mathbf{w} \in \mathbb{R}^d$ such that $\|\mathbf{w}\| > 0$,

$$\lim_{\alpha\to\infty}\frac{\|\nabla F(\alpha\mathbf{w}+\mathbf{w}^*) - \nabla F(\mathbf{w}^*)\|}{L_0\|\alpha\mathbf{w}+\mathbf{w}^*-\mathbf{w}^*\|} = \lim_{\alpha\to\infty}\frac{\|\nabla F(\alpha\mathbf{w}+\mathbf{w}^*)\|}{\alpha L_0\|\mathbf{w}\|} = \lim_{\alpha\to\infty}\frac{k\alpha^{k-2}\|\mathbf{w}\|^{k-1}}{L_0\|\mathbf{w}\|} = \infty.$$

■

The following result demonstrates the difference in worst-case gradient norm scaling that (L_0, L_1)-smooth functions provide, versus the worst-case scaling of functions satisfying Definition 4.

Proposition 29 *For any function satisfying Assumption 2, and any algorithm producing iterates $(\mathbf{w}_s)_{s\ge1}$ satisfying $\|\mathbf{w}_{s+1}-\mathbf{w}_s\| \le \eta \le 1/L_1$ for every $s \ge 1$, the following inequality holds for every $t > t'$:*

$$\|\nabla F(\mathbf{w}_t)\| - (1+\eta L_1)^{t-t'}\|\nabla F(\mathbf{w}_{t'})\| \le ((1+\eta L_1)^{t-t'} - 1)\frac{L_0}{L_1}.$$

Moreover, this inequality is essentially unimprovable, in the sense that there exists a $(0, \mathcal{O}(L_1))$-smooth function and $\eta \le 1/L_1$ such that $\|\nabla F(\mathbf{w}_{T+1})\| = (1+\mathcal{O}(\eta L_1))^{T+1}\|\nabla F(\mathbf{w}_1)\|$ for any $T \ge 1$, and a $(\mathcal{O}(L_0), \mathcal{O}(L_1))$-smooth function such that $\|\nabla F(\mathbf{w}_1)\| = 0$ and $\|\nabla F(\mathbf{w}_{T+1})\| = \mathcal{O}(L_0/L_1)((1+\mathcal{O}(\eta L_1))^T - 1)$. By contrast, any function satisfying Definition 4 satisfies:

$$\|\nabla F(\mathbf{w}_t)\| - c_k\|\nabla F(\mathbf{w}_{t'})\| \le \max\left\{c_k'\eta^{k-1}(t-t')^{k-1}, L_0\eta(t-t')\right\}.$$

Proof We begin by proving the first claim by by induction on $t - t'$. The base case of $t - t' = 1$ holds by definition, since

$$\begin{aligned}\|\nabla F(\mathbf{w}_t)\| - \|\nabla F(\mathbf{w}_{t-1})\| \leq |\|\nabla F(\mathbf{w}_t)\| - \|\nabla F(\mathbf{w}_{t-1})\|| &\leq \|\nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t-1})\| \\ &\leq (L_0 + L_1 \|\nabla F(\mathbf{w}_{t-1})\|) \|\mathbf{w}_t - \mathbf{w}_{t-1}\| \\ &\leq \eta(L_0 + L_1 \|\nabla F(\mathbf{w}_{t-1})\|).\end{aligned}$$

Now, supposing the claim holds for $t - t' = 1, \ldots, s$, we have that:

$$\begin{aligned}\|\nabla F(\mathbf{w}_t)\| &\leq ((1+\eta L_1)^s - 1)\frac{L_0}{L_1} + (1+\eta L_1)^s \|\nabla F(\mathbf{w}_{t-s})\| \\ &\leq ((1+\eta L_1)^s - 1)\frac{L_0}{L_1} + (1+\eta L_1)^s \left(\eta L_0 + (1+\eta L_1)\left\|\nabla F(\mathbf{w}_{t-(s+1)})\right\|\right) \\ &= ((1+\eta L_1)^{s+1} - 1)\frac{L_0}{L_1} + (1+\eta L_1)^{s+1}\left\|\nabla F(\mathbf{w}_{t-(s+1)})\right\|,\end{aligned}$$

where the first inequality follows by applying the induction hypothesis for $t - t' = s$, the second by applying the induction hypothesis for $t - t' = 1$, and the final equality follows by rearranging the prior line. Thus, the inequality holds also at $t - t' = s + 1$, and thus our claim holds by induction.

To see that this inequality is essentially unimprovable, let us consider first consider, for any $L_1 > 0$, the function:

$$F(x) = \exp(L_1 x).$$

Since $F''(x) = L_1^2 \exp(L_1 x) = L_1 F'(x)$, it follows from Proposition 1 that $F(\cdot)$ is $(0, (e-1)L_1)$-smooth. Notice that, if $x_{s+1} - x_s = \eta = 1/(e-1)L_1$, then, taking $x_1 = 0$ and $t \geq 1$,

$$(1 + (e-1)(e^{1/(e-1)} - 1)\eta L_1)^t F'(x_1) = \exp(t/(e-1)) = \exp(\eta L_1 t) = F'(x_{t+1}).$$

Further, for any $L_0, L_1 > 0$, consider the function:

$$F(x) = \frac{L_0}{2}x^2 \exp(L_1 x) - \frac{L_0 x}{L_1}.$$

Clearly,

$$\begin{aligned}F'(x) &= L_0 x \exp(L_1 x) + \frac{L_1 L_0}{2}x^2 \exp(L_1 x) - \frac{L_0}{L_1} \\ F''(x) &= L_0 \exp(L_1 x) + 2L_1 L_0 x \exp(L_1 x) + \frac{L_1^2 L_0 x^2}{2}\exp(L_1 x) \\ &= L_0\left(1 - \frac{L_1^2 x^2}{2}\right)\exp(L_1 x) + 2L_0 + 2L_1 F'(x).\end{aligned}$$

Noting that $F''(x) \leq 2L_0 + 2L_1 F'(x)$ when $|x| \geq \sqrt{2}/L_1$, and $F''(x) \leq L_1(2+\exp(\sqrt{2})) + 2L_1 F'(x)$ otherwise, it follows that F is $(2(2+\exp(\sqrt{2}))L_0, 2(e-1)L_1)$-smooth (by Proposition 1). Therefore,

whenever $\eta = 1/2(e-1)L_1$,

$$\begin{aligned} F(x_{T+1}) &= \frac{L_0}{L_1}\left(\exp\left(\eta L_1 T + \log(L_1\eta T + \frac{L_1^2\eta^2T^2}{2})\right) - 1\right) \\ &= \frac{L_0}{L_1}\left(\exp\left(\mathcal{O}(\eta L_1 T)\right) - 1\right) \\ &= \frac{L_0}{L_1}\left(\exp\left(T\log(1+\mathcal{O}(2(e-1)\eta L_1))\right) - 1\right) \\ &= \mathcal{O}\left(\frac{2(2+\exp(\sqrt{2}))L_0}{2(e-1)L_1}\right)\left((1+\mathcal{O}(2(e-1)\eta L_1))^T - 1\right), \end{aligned}$$

where the first equality follows by rearranging the definition, the second since $\eta L_1 = \Theta(1)$, the third since $2\frac{c}{1+c}(e-1)\eta L_1 = \frac{2c(e-1)\eta L_1}{1+2c(e-1)\eta L_1} \le \log(1+c2(e-1)\eta L_1) \le 2c(e-1)\eta L_1$, and the fourth by rearranging.

The final inequality follows immediately from Fact 16 and Definition 4, which together imply that

$$\|\nabla F(\mathbf{w}_t)\| - c_k\|\nabla F(\mathbf{w}_{t'})\| \le \max\left\{c_k'\eta^{k-1}(t-t')^{k-1}, L_0\eta(t-t')\right\}.$$

■

C.2. Bounding $\mathrm{comp}(\tau)$ from Lemma 8

In order to use Theorem 26, recall that we must be able to bound the quantity $\mathrm{comp}(\tau_{T+1}(\delta))$. To accomplish this, we show that, if one can find "good" times t' near to the "bad" time t (that is, $t-t'$ is "small"), then it is possible to bound $\mathrm{comp}(\tau)$. We remark that this result generalizes the compensation argument of (Faw et al., 2022) to functions satisfying Definition 4.

Lemma 30 *Suppose that $F(\cdot)$ satisfies Definition 4 for some constants $k \ge 2, c_k \ge 1, c_k' > 0$. Fix any time $t \in [T]$, and let $S_{[t]}^{\mathrm{comp}} \subset [T]$ be any set such that $t > \max(S_{[t]}^{\mathrm{comp}})$ and $|S_{[t]}^{\mathrm{comp}}| \le n_{\mathrm{comp}} := \left\lceil \frac{4c_k^3(\sigma_1 - (1-\varepsilon-\varepsilon'))_+}{\varepsilon'''} \right\rceil$ (where $(x)_+ := \max\{0,x\}$). Then, assuming $\{\mathbf{w}_t\}_{t\ge1}$ are the iterates corresponding to (AG-Norm), we have that either $|S_{[t]}^{\mathrm{comp}}| < n_{\mathrm{comp}}$, or:*

$$\begin{aligned} &(\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t'\in S_{[t]}^{\mathrm{comp}}} \varepsilon'''\tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 \\ &\le \frac{\varepsilon'''\eta n_{\mathrm{comp}}}{2c_k^2}\max\left\{c_k'\eta^{k-1}, L_0\eta\right\}(t - \min(S_{[t]}^{\mathrm{comp}}))^{k-1}. \end{aligned}$$

Proof We first show that

$$\frac{\tilde{\eta}_t}{4c_k^3}\|\nabla F(\mathbf{w}_t)\|^2 - \tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 \le \frac{\eta}{2c_k^2}\max\left\{c_k'\|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0\|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}. \quad (17)$$

To see this, first observe that, recalling the definition of $\tilde{\eta}_t$ from Definition 2,

$$\begin{aligned}
\frac{\frac{\tilde{\eta}_t}{c_k} - \tilde{\eta}_{t'}}{\eta} &= \frac{1}{c_k \widetilde{b}_t} - \frac{1}{\widetilde{b}_{t'}} \\
&= \frac{\widetilde{b}_{t'}^2 - c_k^2 \widetilde{b}_t^2}{c_k \widetilde{b}_t \widetilde{b}_{t'} (c_k \widetilde{b}_t + \widetilde{b}_{t'})} \\
&= \frac{b_{t'-1}^2 - c_k^2 b_{t-1}^2 + \|\nabla F(\mathbf{w}_{t'})\|^2 - c_k^2 \|\nabla F(\mathbf{w}_t)\|^2}{c_k \widetilde{b}_t \widetilde{b}_{t'} (c_k \widetilde{b}_t + \widetilde{b}_{t'})} \\
&\leq \frac{\|\nabla F(\mathbf{w}_{t'})\|^2 - c_k^2 \|\nabla F(\mathbf{w}_t)\|^2}{c_k \widetilde{b}_t \widetilde{b}_{t'} (c_k \widetilde{b}_t + \widetilde{b}_{t'})} \\
&= \frac{(\|\nabla F(\mathbf{w}_{t'})\| - c_k \|\nabla F(\mathbf{w}_t)\|)(\|\nabla F(\mathbf{w}_{t'})\| + c_k \|\nabla F(\mathbf{w}_t)\|)}{c_k \widetilde{b}_t \widetilde{b}_{t'} (c_k \widetilde{b}_t + \widetilde{b}_{t'})} \\
&\leq \frac{\max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\} (\|\nabla F(\mathbf{w}_{t'})\| + c_k \|\nabla F(\mathbf{w}_t)\|)}{c_k \widetilde{b}_t \widetilde{b}_{t'} (c_k \widetilde{b}_t + \widetilde{b}_{t'})} \\
&\leq \frac{\max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}}{c_k \|\nabla F(\mathbf{w}_t)\| \|\nabla F(\mathbf{w}_{t'})\|}
\end{aligned}$$

where we use the fact that $c_k \geq 1$ and $b_{t-1}^2 \geq b_{t'-1}^2$ (since $t \geq t'$) for the first inequality, the definition of Definition 4 for the second, and the definition of $\widetilde{b}_t$ from Definition 2 for the third. Now, either $\|\nabla F(\mathbf{w}_t)\| \geq 2 \max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}$, or not. In the first case, we note that

$$\begin{aligned}
c_k \|\nabla F(\mathbf{w}_{t'})\| &\geq \|\nabla F(\mathbf{w}_t)\| - \max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\} \\
&\geq \|\nabla F(\mathbf{w}_t)\| - \frac{1}{2}\|\nabla F(\mathbf{w}_t)\| = \frac{1}{2}\|\nabla F(\mathbf{w}_t)\|,
\end{aligned}$$

from which we may conclude that

$$\begin{aligned}
\frac{\tilde{\eta}_t}{4c_k^3}\|\nabla F(\mathbf{w}_t)\|^2 - \tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 &\leq \frac{1}{4c_k^2}\left(\frac{\tilde{\eta}_t}{c_k} - \tilde{\eta}_{t'}\right)\|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \frac{\eta \max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}}{4c_k^3 \|\nabla F(\mathbf{w}_t)\| \|\nabla F(\mathbf{w}_{t'})\|}\|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \frac{\eta \max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}}{4c_k^3 \|\nabla F(\mathbf{w}_t)\| \frac{1}{2c_k}\|\nabla F(\mathbf{w}_t)\|}\|\nabla F(\mathbf{w}_t)\|^2 \\
&= \frac{\eta \max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}}{2c_k^2}.
\end{aligned}$$

In the alternate case that $\|\nabla F(\mathbf{w}_t)\| < 2\max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0 \|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}$, we obtain:

$$\begin{aligned}
\frac{\tilde{\eta}_t}{4c_k^3}\|\nabla F(\mathbf{w}_t)\|^2 - \tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 &\le \frac{\tilde{\eta}_t}{4c_k^3}\|\nabla F(\mathbf{w}_t)\|^2 \\
&\le \frac{\eta}{4c_k^3}\|\nabla F(\mathbf{w}_t)\| \\
&< \frac{\eta}{4c_k^3} 2\max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0\|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\} \\
&= \frac{\eta}{2c_k^3}\max\left\{c_k' \|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0\|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\},
\end{aligned}$$

which, since $c_k \ge 1$, establishes (17). The lemma follows straightforwardly from (17). Indeed, note that the claimed inequality is trivially true whenever $|S_{[t]}^{\mathrm{comp}}| = n_{\mathrm{comp}} = 0$, since this implies that $\sigma_1 \le 1 - \varepsilon - \varepsilon'$. Otherwise, when $n_{\mathrm{comp}} > 0$, we have that

$$\begin{aligned}
&(\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t'\in S_{[t]}^{\mathrm{comp}}} \varepsilon'''\tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 \\
&= \sum_{t'\in S_{[t]}^{\mathrm{comp}}} \frac{\sigma_1 - (1-\varepsilon-\varepsilon')}{n_{\mathrm{comp}}}\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2 - \varepsilon'''\tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 \\
&\le \varepsilon''' \sum_{t'\in S_{[t]}^{\mathrm{comp}}} \frac{\tilde{\eta}_t}{4c_k^2}\|\nabla F(\mathbf{w}_t)\|^2 - \tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 \\
&\le \varepsilon''' \sum_{t'\in S_{[t]}^{\mathrm{comp}}} \frac{\eta\max\left\{c_k'\|\mathbf{w}_t - \mathbf{w}_{t'}\|^{k-1}, L_0\|\mathbf{w}_t - \mathbf{w}_{t'}\|\right\}}{2c_k^2}.
\end{aligned}$$

Thus, by Fact 16, we conclude that:

$$\begin{aligned}
&(\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_t\|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t'\in S_{[t]}^{\mathrm{comp}}} \varepsilon'''\tilde{\eta}_{t'}\|\nabla F(\mathbf{w}_{t'})\|^2 \\
&\le n_{\mathrm{comp}}\varepsilon''' \frac{\eta\max\left\{c_k'\eta^{k-1}(t - \min(S_{[t]}^{\mathrm{comp}}))^{k-1}, L_0\eta(t-\min(S_{[t]}^{\mathrm{comp}}))\right\}}{2c_k^2} \\
&\le n_{\mathrm{comp}}\varepsilon''' \frac{\eta\max\left\{c_k'\eta^{k-1}, L_0\eta\right\}}{2c_k^2}(t - \min(S_{[t]}^{\mathrm{comp}}))^{k-1}
\end{aligned}$$

as claimed. ■

We now show how to translate Lemma 30 directly into a bound on $\mathrm{comp}(\tau)$. This shows that, in order to bound $\mathrm{comp}(\tau)$, it suffices to bound $\mathbb{E}\left[|S_{\mathrm{good}}(\tau_{T+1}(\delta))^c|^k\right]$ by a "sufficiently small" quantity (say, $\mathcal{O}(\log(T))$).

Lemma 11 *Suppose that $F(\cdot)$ satisfies Definition 4 for some constants $k \ge 2$, $c_k \ge 1$, and $c_k' > 0$. Let $\tau \in [2, T+1]$ be any (possibly random) time. Then, recalling $\mathrm{comp}(\tau)$ and $S^{\mathrm{comp}}(\tau)$*

from Lemma 8, there is an explicit construction of $S^{\text{comp}}(\tau)$ (the subset of "good" times used to compensate for $S_{\text{good}}(\tau)^c$) such that, for any $\varepsilon, \varepsilon', \varepsilon''' \in (0,1)$ such that $\varepsilon + \varepsilon' < 1$ and $n_{\text{comp}} = \lceil 4c_k^3(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ / \varepsilon''' \rceil$ (and taking $(x)_+ := \max\{0, x\}$) $\text{comp}(\tau)$ can be bounded as follows:

$$\text{comp}(\tau) \leq \eta(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ c_k \|\nabla F(\mathbf{w}_1)\| \, \mathbb{E}\left[|S_{\text{good}}(\tau)^c|\right] + \eta n_{\text{comp}}^{k-1} \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} \left((\sigma_1 - (1-\varepsilon-\varepsilon'))_+ + \frac{\varepsilon''' n_{\text{comp}}}{2c_k^3}\right) \mathbb{E}\left[|S_{\text{good}}(\tau)^c|^k\right].$$

Proof First, let us construct $\widetilde{S}(\tau)$ in the same manner as in (Faw et al., 2022, Lemma 11). In particular, denote $\tau_{[i]}^{\text{bad}}$ as the ith largest "bad" time in $S_{\text{good}}(\tau)^c$, i.e., $\tau_{[1]}^{\text{bad}} = \max(S_{\text{good}}(\tau)^c)$, and, for every $i \in [2, |S_{\text{good}}(\tau)^c|]$,

$$\tau_{[i]}^{\text{bad}} = \max\left\{t \in S_{\text{good}}(\tau)^c : t < \tau_{[i-1]}^{\text{bad}}\right\}.$$

Then, to every "bad" time $\tau_{[i]}^{\text{bad}}$, associate a set $S_{[i]}^{\text{comp}}$ of the largest (at most) $n_{\text{comp}} = \max\left\{0, \left\lceil \frac{4c_k^3(\sigma_1 - (1-\varepsilon-\varepsilon'))}{\varepsilon'''} \right\rceil\right\}$ "good" times before $\tau_{[i]}^{\text{bad}}$ that are not assigned to another $\tau_{[i']}^{\text{bad}} > \tau_{[i]}^{\text{bad}}$. That is, denoting

$$\begin{aligned}
\tau_{[0,n_{\text{comp}}]}^{\text{good}} &:= +\infty \\
\tau_{[i,1]}^{\text{good}} &:= \max\left\{S_{\text{good}}(\tau) \cap [1, \min\left\{\tau_{[i]}^{\text{bad}}, \tau_{[i-1,n_{\text{comp}}]}^{\text{good}}\right\})\right\} \\
\tau_{[i,j+1]}^{\text{good}} &:= \max\left\{t \in S_{\text{good}}(\tau) : t < \min\left\{\tau_{[i]}^{\text{bad}}, \tau_{[i,j]}^{\text{good}}\right\}\right\},
\end{aligned}$$

where, when the maximum does not exist, we take $\tau_{[i,j]}^{\text{good}} = -\infty$. We can then take

$$S_{[i]}^{\text{comp}} := \left\{\tau_{[i,j]}^{\text{good}} : j \in [n_{\text{comp}}], \tau_{[i,j]}^{\text{good}} > -\infty\right\}.$$

Then, by (Faw et al., 2022, Lemma 11), we have that, for some index $i^* \in [|S_{\text{good}}(\tau)^c|]$, and for every $i < i^*$,

$$|S_{[i]}^{\text{comp}}| = n_{\text{comp}} \text{ and } \tau_{[i]}^{\text{bad}} - \min(S_{[i]}^{\text{comp}}) \leq n_{\text{comp}} |S_{\text{good}}(\tau)^c| \text{ if } n_{\text{comp}} > 0. \tag{18}$$

For the remaining $i \geq i^*$, $\tau_{[i]}^{\text{bad}} \leq \tau_{[i^*]}^{\text{bad}} \leq n_{\text{comp}} |S_{\text{good}}(\tau)^c|$. Finally, we take:

$$S^{\text{comp}}(\tau) = \cup_{i \in [|S_{\text{good}}(\tau)^c|]} S_{[i]}^{\text{comp}} \quad \text{and} \quad \widetilde{S}(\tau) = S_{\text{good}}(\tau) \setminus S^{\text{comp}}(\tau)$$

We use these compensation sets to bound the quantity $\text{comp}(\tau)$ from Lemma 8. Indeed, we can decompose this quantity as follows:

$$\begin{aligned}
\text{comp}(\tau) &= \mathbb{E}\left[\sum_{t \in S_{\text{good}}(\tau)^c} (\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_t \|\nabla F(\mathbf{w}_t)\|^2 - \sum_{t' \in S^{\text{comp}}(\tau)} \varepsilon''' \tilde{\eta}_{t'} \|\nabla F(\mathbf{w}_{t'})\|^2\right] \\
&= \mathbb{E}\left[\sum_{i < i^*} (\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_{\tau_{[i]}^{\text{bad}}} \left\|\nabla F(\mathbf{w}_{\tau_{[i]}^{\text{bad}}})\right\|^2 - \sum_{t' \in S_{[i]}^{\text{comp}}} \varepsilon''' \tilde{\eta}_{t'} \|\nabla F(\mathbf{w}_{t'})\|^2\right] \\
&\quad + \mathbb{E}\left[\sum_{i \geq i^*} (\sigma_1 - (1-\varepsilon-\varepsilon'))\tilde{\eta}_{\tau_{[i]}^{\text{bad}}} \left\|\nabla F(\mathbf{w}_{\tau_{[i]}^{\text{bad}}})\right\|^2 - \sum_{t' \in S_{[i]}^{\text{comp}}} \varepsilon''' \tilde{\eta}_{t'} \|\nabla F(\mathbf{w}_{t'})\|^2\right],
\end{aligned}$$

To obtain a bound on the first term, we can use Lemma 30. For the second, we trivially lower-bound $\sum_{t' \in S^{\text{comp}}_{[i]}} \varepsilon''' \tilde{\eta}_{t'} \left\|\nabla F(\mathbf{w}_{t'})\right\|^2 \geq 0$. The resulting bound is:

$$\begin{aligned}\text{comp}(\tau) \leq \frac{\varepsilon''' \eta \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} n_{\text{comp}}}{2c_k^2} \mathbb{E}\left[\sum_{i<i^*} (\tau^{\text{bad}}_{[i]} - \min(S^{\text{comp}}_{[i]}))^{k-1}\right] \\ + (\sigma_1 - (1-\varepsilon-\varepsilon'))\mathbb{E}\left[\sum_{i \geq i^*} \tilde{\eta}_{\tau^{\text{bad}}_{[i]}} \left\|\nabla F(\mathbf{w}_{\tau^{\text{bad}}_{[i]}})\right\|^2\right].\end{aligned}$$

Next, using (18) to bound $\tau^{\text{bad}}_{[i]} - \min(S^{\text{comp}}_{[i]})$ for each $i < i^*$, and recalling $\tilde{\eta}_t \left\|\nabla F(\mathbf{w}_t)\right\|^2 \leq \eta \left\|\nabla F(\mathbf{w}_t)\right\|$ for every t, the above bound becomes:

$$\begin{aligned}\text{comp}(\tau) \leq \frac{\varepsilon''' \eta \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} n_{\text{comp}}^k}{2c_k^3} \mathbb{E}\left[\sum_{i<i^*} |S_{\text{good}}(\tau)^c|^{k-1}\right] \\ + (\sigma_1 - (1-\varepsilon-\varepsilon'))_+ \mathbb{E}\left[\sum_{i \geq i^*} \eta \left\|\nabla F(\mathbf{w}_{\tau^{\text{bad}}_{[i]}})\right\|\right].\end{aligned}$$

Now, notice that both summation ranges $i < i^*$ and $i \geq i^*$ are of size at most $|S_{\text{good}}(\tau)^c|$. Thus, the first term can be bounded as:

$$\mathbb{E}\left[\sum_{i<i^*} |S_{\text{good}}(\tau)^c|^{k-1}\right] \leq \mathbb{E}\left[|S_{\text{good}}(\tau)^c|^k\right].$$

To bound the second term, we apply Definition 4 and Fact 16, together with the above construction, to obtain:

$$\begin{aligned}\mathbb{E}\left[\sum_{i \geq i^*} \left\|\nabla F(\mathbf{w}_{\tau^{\text{bad}}_{[i]}})\right\|\right] &\leq \mathbb{E}\left[\sum_{i \geq i^*} c_k \left\|\nabla F(\mathbf{w}_1)\right\| + \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} (\tau^{\text{bad}}_{[i]})^{k-1}\right] \\ &\leq \mathbb{E}\left[\sum_{i \geq i^*} c_k \left\|\nabla F(\mathbf{w}_1)\right\| + \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} n_{\text{comp}}^{k-1} |S_{\text{good}}(\tau)^c|^{k-1}\right] \\ &\leq c_k \left\|\nabla F(\mathbf{w}_1)\right\| \mathbb{E}\left[|S_{\text{good}}(\tau)^c|\right] + \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} n_{\text{comp}}^{k-1} \mathbb{E}\left[|S_{\text{good}}(\tau)^c|^k\right]\end{aligned}$$

Collecting results, we have that:

$$\begin{aligned}\text{comp}(\tau) \leq \eta(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ c_k \left\|\nabla F(\mathbf{w}_1)\right\| \mathbb{E}\left[|S_{\text{good}}(\tau)^c|\right] \\ + \eta n_{\text{comp}}^{k-1} \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} \left((\sigma_1 - (1-\varepsilon-\varepsilon'))_+ + \frac{\varepsilon''' n_{\text{comp}}}{2c_k^3}\right) \mathbb{E}\left[|S_{\text{good}}(\tau)^c|^k\right],\end{aligned}$$

as claimed. ■

The next result, combined with Lemma 11, completes our goal of bounding $\text{comp}(\tau_{T+1}(\delta))$ by $\text{poly}\log(T)$.

Lemma 12 *Let $\tau_{T+1}(\delta) \leq T+1$ be the stopping time with respect to $(\mathcal{F}_{s-1})_{s\geq 1}$ from Definition 9. Recall the set $S_{\text{good}}(\tau_{T+1}(\delta))$ from Definition 21, and denote $S_{\text{good}}(\tau_{T+1}(\delta))^c = [\tau_{T+1}(\delta)-1] \setminus S_{\text{good}}(\tau_{T+1}(\delta))$. Let $f(T) = e + \frac{e\sigma_0^2(T-1)+e(1+\sigma_1^2+c_{\text{L}})\mathbb{E}\left[\sum_{t<\tau_T(\delta)}\|\nabla F(\mathbf{w}_t)\|^2\right]}{b_0^2\delta}$. Then, for any $k \geq 1$, the iterates of (AG-Norm) satisfy (under Assumption 4):*

$$\mathbb{E}\left[|S_{\text{good}}(\tau_{T+1}(\delta))^c|^k\right] \leq \left(\frac{(k+1)\sigma_1^2\log(f(T))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^k.$$

Proof Note that we can write $|S_{\text{good}}(\tau_{T+1}(\delta))^c|$ as:

$$|S_{\text{good}}(\tau_{T+1}(\delta))^c| = \sum_{t<\tau_{T+1}(\delta)} \mathbb{1}\{t \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}.$$

Thus, by the Multinomial theorem, we have that

$$\begin{aligned}
|S_{\text{good}}(\tau_{T+1}(\delta))^c|^k &= \sum_{\substack{k_1,\ldots,k_{\tau_{T+1}(\delta)-1}\geq 0 \\ k_1+\ldots+k_{\tau_{T+1}(\delta)-1}=k}} \binom{k}{k_1,\ldots,k_{\tau_{T+1}(\delta)-1}} \prod_{t<\tau_{T+1}(\delta)} \mathbb{1}\{t \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}^{k_t} \\
&= \sum_{s=1}^{k} \sum_{1\leq t_1<\ldots<t_s\leq\tau_{T+1}(\delta)-1} \sum_{\substack{k_{t_1},\ldots,k_{t_s}>0 \\ k_{t_1}+\ldots+k_{t_s}=k}} \binom{k}{k_{t_1},\ldots,k_{t_s}} \prod_{\ell\in[s]} \mathbb{1}\{t_\ell \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}^{k_{t_\ell}} \\
&= \sum_{s=1}^{k} \sum_{1\leq t_1<\ldots<t_s\leq\tau_{T+1}(\delta)-1} \sum_{\substack{k_{t_1},\ldots,k_{t_s}>0 \\ k_{t_1}+\ldots+k_{t_s}=k}} \binom{k}{k_{t_1},\ldots,k_{t_s}} \prod_{\ell\in[s]} \mathbb{1}\{t_\ell \in S_{\text{good}}(\tau_{T+1}(\delta))^c\} \\
&= \sum_{s=1}^{k} \sum_{1\leq t_1<\ldots<t_s\leq\tau_{T+1}(\delta)-1} \prod_{\ell\in[s]} \mathbb{1}\{t_\ell \in S_{\text{good}}(\tau_{T+1}(\delta))^c\} \sum_{\substack{k_{t_1},\ldots,k_{t_s}>0 \\ k_{t_1}+\ldots+k_{t_s}=k}} \binom{k}{k_{t_1},\ldots,k_{t_s}},
\end{aligned}$$

where in the second line, we rewrite the first summation as a sum over all possible support sets $\{t_1,\ldots,t_s\} \subset [\tau_{T+1}(\delta)-1]$ of size $s \in [k]$ of terms included in the summation. The third equality follows immediately from the second, since each $k_\ell > 0$. The final equality follows by rearranging the terms in the prior one. Now, by another application of the Multinomial theorem, we have that

$$\sum_{\substack{k_{t_1},\ldots,k_{t_s}>0 \\ k_{t_1}+\ldots+k_{t_s}=k}} \binom{k}{k_{t_1},\ldots,k_{t_s}} \leq \sum_{\substack{k_{t_1},\ldots,k_{t_s}\geq 0 \\ k_{t_1}+\ldots+k_{t_s}=k}} \binom{k}{k_{t_1},\ldots,k_{t_s}} = s^k.$$

Combining this with the above, we have the following:

$$|S_{\text{good}}(\tau_{T+1}(\delta))^c|^k \leq \sum_{s=1}^{k} s^k \sum_{1\leq t_1<\ldots<t_s\leq\tau_{T+1}(\delta)-1} \prod_{\ell\in[s]} \mathbb{1}\{t_\ell \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}.$$

We claim that, for any $s \geq 1$, the inner summation term above is bounded in expectation by:

$$\mathbb{E}\left[\sum_{1\leq t_1<\ldots<t_s\leq\tau_{T+1}(\delta)-1}\prod_{\ell\in[s]}\mathbb{1}\{t_\ell\in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\right] \leq \left(\frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log\left(f(T)\right)\right)^s, \tag{19}$$

where $f(T) = e + \frac{e\sigma_0^2(T-1)+e(1+\sigma_1^2+c_{\text{L}})\mathbb{E}\left[\sum_{t<\tau_T(\delta)}\|\nabla F(\mathbf{w}_t)\|^2\right]}{\delta b_0^2}$. We prove (19) via induction on s. We begin by observing that, for any $t' \geq 0$,

$$\mathbb{E}\left[\sum_{t=t'+1}^{\tau_{T+1}(\delta)-1}\mathbb{1}\{t\in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\mid\mathcal{F}_{t'-1}\right] \leq \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log\left(f(T)\right). \tag{20}$$

To see this, first note that, by Definition 21, and since $\{t<\tau_{T+1}(\delta)\}\in\mathcal{F}_{t-1}$ by Lemma 23, for any $t'\geq 0$,

$$\begin{aligned}
&\frac{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2)}{\sigma_1^2}\mathbb{E}\left[\sum_{t=t'+1}^{T}\mathbb{1}\{t\in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\mid\mathcal{F}_{t'-1}\right]\\
&\leq\sum_{t=t'+1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mid\mathcal{F}_{t-1}\right]\mathbb{1}\{t\in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\mid\mathcal{F}_{t'-1}\right]\\
&\leq\sum_{t=t'+1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mid\mathcal{F}_{t-1}\right]\mathbb{1}\{t<\tau_{T+1}(\delta)\}\mid\mathcal{F}_{t'-1}\right]\\
&=\sum_{t=t'+1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mathbb{1}\{t<\tau_{T+1}(\delta)\}\mid\mathcal{F}_{t-1}\right]\mid\mathcal{F}_{t'-1}\right]\\
&=\mathbb{E}\left[\sum_{t=t'+1}^{\tau_{T+1}(\delta)-1}\frac{\|\boldsymbol{g}_t\|^2}{b_t^2}\mid\mathcal{F}_{t'-1}\right].
\end{aligned}$$

Now, by Lemma 15, we have that

$$\begin{aligned}
\sum_{t=t'+1}^{\tau_{T+1}(\delta)-1}\frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \leq \sum_{t=t'+1}^{\tau_{T+1}(\delta)-1}\frac{\|\boldsymbol{g}_t\|^2}{b_0^2+\sum_{s=t'+1}^{t}\|\boldsymbol{g}_s\|^2} &\leq 1+\sum_{t=t'+1}^{\tau_{T+1}(\delta)-2}\frac{\|\boldsymbol{g}_t\|^2}{b_0^2+\sum_{s=t'+1}^{t}\|\boldsymbol{g}_t\|^2}\\
&\leq 1+\log\left(\frac{b_0^2+\sum_{t=t'+1}^{\tau_{T+1}(\delta)-2}\|\boldsymbol{g}_t\|^2}{b_0^2}\right).
\end{aligned}$$

Now, by Items 4, 5 and 6 of Lemma 23, we have that, almost surely,

$$\begin{aligned}
\sum_{t=t'+1}^{\tau_{T+1}(\delta)-2}\|\boldsymbol{g}_t\|^2 &\leq \sum_{t<\tau_{T+1}(\delta)-1}\|\boldsymbol{g}_t\|^2+c_{\text{L}}\|\nabla F(\mathbf{w}_t)\|^2 = \sum_{t<\tau_{\tau_{T+1}(\delta)-1}(\delta)}\|\boldsymbol{g}_t\|^2+c_{\text{L}}\|\nabla F(\mathbf{w}_t)\|^2\\
&= S_{\tau_{T+1}(\delta)-1} \leq \frac{\mathbb{E}\left[S_T\right]}{\delta} \leq \frac{(T-1)\sigma_0^2+(1+\sigma_1^2+c_{\text{L}})\mathbb{E}\left[\sum_{t<\tau_T(\delta)}\|\nabla F(\mathbf{w}_t)\|^2\right]}{\delta}.
\end{aligned}$$

Therefore, collecting these results, we conclude that, for any $t' \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t'+1}^{\tau_{T+1}(\delta)-1} \mathbb{1}\{t \in S_{\text{good}}(\tau_{T+1}(\delta))^c\} \mid \mathcal{F}_{t'-1}\right] \\
&\leq \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\mathbb{E}\left[\sum_{t=t'+1}^{\tau_{T+1}(\delta)-1} \frac{\|\boldsymbol{g}_t\|^2}{b_t^2} \mid \mathcal{F}_{t'-1}\right] \\
&\leq \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\mathbb{E}\left[1+\log\left(1+\frac{(T-1)\sigma_0^2+(1+\sigma_1^2+c_{\text{L}})\mathbb{E}\left[\sum_{t<\tau_T(\delta)}\|\nabla F(\mathbf{w}_t)\|^2\right]}{\delta b_0^2}\right) \mid \mathcal{F}_{t'-1}\right] \\
&= \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log\left(f(T)\right),
\end{aligned}$$

as claimed.

Now, the base case of $s = 1$ for (19) follows immediately from (20) with $t' = 0$. Let us now suppose that the claim (19) holds for some $s \geq 1$. Then, to apply the induction hypothesis, we begin by decomposing:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{1\leq t_1<\ldots<t_{s+1}\leq\tau_{T+1}(\delta)-1}\prod_{\ell\in[s+1]}\mathbb{1}\{t_\ell \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\right] \\
&= \sum_{1\leq t_1<\ldots<t_s\leq T}\mathbb{E}\left[\mathbb{1}\{t_i \in S_{\text{good}}(\tau_{T+1}(\delta))^c\ \forall i\in[s]\}\sum_{t_{s+1}=t_s+1}^{\tau_{T+1}(\delta)-1}\mathbb{1}\{t_{s+1} \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\right].
\end{aligned}$$

Notice that the above expectation is a product of two terms: indicators depending of times $t_1, \ldots, t_s$, and those depending on $t_{s+1} > t_s$. Therefore, since, by Lemma 23 and Definition 21,

$$\{t \in S_{\text{good}}(\tau_{T+1}(\delta))^c\} = \{t < \tau_{T+1}(\delta)\} \cap \{t \text{ is ``good''}\} \in \mathcal{F}_{t_s-1},$$

we may apply the tower rule of expectations and the inequality from (20):

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{t_i \in S_{\text{good}}(\tau_{T+1}(\delta))^c\ \forall i\in[s]\}\sum_{t_{s+1}=t_s+1}^{\tau_{T+1}(\delta)-1}\mathbb{1}\{t_{s+1} \in S_{\text{good}}(\tau_{T+1}(\delta))^c\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{t_i \in S_{\text{good}}(\tau_{T+1}(\delta))^c\ \forall i\in[s]\}\sum_{t_{s+1}=t_s+1}^{\tau_{T+1}(\delta)-1}\mathbb{1}\{t_{s+1} \in S_{\text{good}}(\tau_{T+1}(\delta))^c\} \mid \mathcal{F}_{t_s-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}\{t_i \in S_{\text{good}}(\tau_{T+1}(\delta))^c\ \forall i\in[s]\}\mathbb{E}\left[\sum_{t_{s+1}=t_s+1}^{\tau_{T+1}(\delta)-1}\mathbb{1}\{t_{s+1} \in S_{\text{good}}(\tau_{T+1}(\delta))^c\} \mid \mathcal{F}_{t_s-1}\right]\right] \\
&\leq \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log(f(T))\mathbb{E}\left[\mathbb{1}\{t_i \in S_{\text{good}}(\tau_{T+1}(\delta))^c\ \forall i\in[s]\}\right].
\end{aligned}$$

Therefore, summing the above expression over $1 \leq t_1 < \ldots < t_s \leq T$ and applying the induction hypothesis, we conclude that:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{1\leq t_1<\ldots<t_{s+1}\leq \tau_{T+1}(\delta)-1}\prod_{\ell\in[s+1]}\mathbb{1}\{t_\ell \in S_{\mathrm{good}}(\tau_{T+1}(\delta))^c\}\right] \\
&\leq \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log(f(T))\mathbb{E}\left[\sum_{1\leq t_1<\ldots<t_s\leq \tau_{T+1}(\delta)-1}\mathbb{1}\{t_i \in S_{\mathrm{good}}(\tau_{T+1}(\delta))^c\ \forall i\in[s]\}\right] \\
&= \frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log(f(T))\mathbb{E}\left[\sum_{1\leq t_1<\ldots<t_s\leq \tau_{T+1}(\delta)-1}\prod_{\ell\in[s]}\mathbb{1}\{t_\ell \in S_{\mathrm{good}}(\tau_{T+1}(\delta))^c\}\right] \\
&\leq \left(\frac{\sigma_1^2}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\log(f(T))\right)^{s+1},
\end{aligned}$$

which establishes (19) by induction.

Finally, using (19), we conclude that

$$\begin{aligned}
\mathbb{E}\left[|S_{\mathrm{good}}(\tau_{T+1}(\delta))^c|^k\right] &\leq \sum_{s\in[k]} s^k \mathbb{E}\left[\sum_{1\leq t_1<\ldots<t_s\leq T}\prod_{\ell\in[s]}\mathbb{1}\{t_\ell \notin S_{\mathrm{good}}(\tau_{T+1}(\delta))\}\right] \\
&\leq \sum_{s\in[k]} s^k \left(\frac{\sigma_1^2\log(f(T))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^s.
\end{aligned}$$

Now, finally noting that, for any $x \geq 1$,

$$\sum_{s\in[k]} s^k x^s \leq x^k \sum_{s\in[k]} s^k \leq x^k \int_1^{k+1} s^k = \frac{x^k((k+1)^{k+1}-1)}{k+1} \leq x^k(k+1)^k,$$

we conclude that

$$\mathbb{E}\left[|S_{\mathrm{good}}(\tau_{T+1}(\delta))^c|^k\right] \leq \left(\frac{(k+1)\sigma_1^2\log(f(T))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^k,$$

as claimed. ■

C.3. Bounding the sum of "bad" gradients by the sum of "good" ones

We recall from Theorem 26 that, in order to use this bound, we need to show that the sum of "bad" gradients can be upper-bounded (relatively) by the sum of "good" ones. It turns out, for functions satisfying Definition 4, this is possible, as we now show.

Lemma 31 *Let $\tau \geq 1$ be any (possibly random) time, and consider any (possibly random) set $S(\tau) \subseteq [\tau-1]$. Denote $S(\tau)^c = [\tau-1] \setminus S(\tau)$. Then, assuming $F(\cdot)$ satisfies Definition 4, the*

following is satisfied deterministically:

$$\sum_{t\in S(\tau)^c} \|\nabla F(\mathbf{w}_t)\|^2 \leq 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}|S(\tau)^c|^{2k-1} + 2c_k^2\|\nabla F(\mathbf{w}_1)\|^2|S(\tau)^c| + 2c_k^2\sum_{t\in S(\tau)}\|\nabla F(\mathbf{w}_t)\|^2.$$

In particular, recalling $\tau_{T+1}(\delta)$ as the stopping time from Definition 9 and $S_{\text{good}}(\tau_{T+1}(\delta))$ the set of "good" times before $\tau_{T+1}(\delta)$ from Definition 21, we have that, for any $\widetilde{S}(\tau_{T+1}(\delta)) \subseteq S_{\text{good}}(\tau_{T+1}(\delta))$ such that $\mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right] \leq (1+n_{\text{comp}})\mathbb{E}\left[|S_{\text{good}}(\tau_{T+1}(\delta))^c|\right]$, we have that:

$$\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))^c}\|\nabla F(\mathbf{w}_t)\|^2\right] \leq c_{\text{B1}} + c_{\text{B2}}\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))}\|\nabla F(\mathbf{w}_t)\|^2\right],$$

where

$$c_{\text{B1}} = 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}(n_{\text{comp}}+1)^{2k-1}\left(\frac{2k\sigma_1^2\log(f(\tau_{T+1}(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^{2k-1} + 2c_k^2\|\nabla F(\mathbf{w}_1)\|^2(n_{\text{comp}}+1)\left(\frac{2\sigma_1^2\log(f(\tau_{T+1}(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right),$$

$$c_{\text{B2}} = 2c_k^2$$

Proof The proof of this result follows a similar argument as used in Lemma 11. The main idea here is to, for every $t \in S(\tau)^c$ in decreasing order, find the first available time $t' \in S(\tau)$ which has not been associated with an earlier time from $S(\tau)^c$. Then, using Definition 4, we show that, as long as t and t' are not too far apart, then $\|\nabla F(\mathbf{w}_t)\|^2$ and $\|\nabla F(\mathbf{w}_{t'})\|^2$ must also be close. For some times $t \in S(\tau)^c$, there may not be such a $t' \in S(\tau)$. However, because of the greedy construction, these times must be relatively small (roughly within the first $|S(\tau)^c|$ time steps). Thus, as long as $|S(\tau)^c|$ is not "too big" (in expectation), then we can still bound these remaining terms. We now make these arguments precise.

To begin, note that for every $t, t' \geq 1$, by Definition 4,

$$\|\nabla F(\mathbf{w}_t)\|^2 \leq 2c_k^2\|\nabla F(\mathbf{w}_{t'})\|^2 + 2c_k'^2\|\mathbf{w}_t - \mathbf{w}_{t'}\|^{2(k-1)}.$$

We use this bound as follows: let us index the times in $S(\tau)^c$, denoting $\widetilde{\tau}_{[i]}$ to be the ith largest time in $S(\tau)^c$, i.e.,

$$\widetilde{\tau}_{[1]} = \max(S(\tau)^c) \quad \text{and} \quad \widetilde{\tau}_{[i]} = \max(S(\tau)^c \setminus \left(\cup_{i'=1}^{i-1}\left\{\widetilde{\tau}_{[i']}\right\}\right)) \; \forall i \in [2, |S(\tau)^c|].$$

To each $\widetilde{\tau}_{[i]}$ in decreasing order of time, associate the largest time $\widetilde{\tau}_{[i]}^{\text{good}}$ in $S(\tau)$ before $\widetilde{\tau}_{[i]}$ which has not already been associated with some other $\widetilde{\tau}_{[i']} > \widetilde{\tau}_{[i]}$, as long as such a time exists. In particular, we take

$$\widetilde{\tau}_{[i]}^{\text{good}} = \max\left\{t \in S(\tau) : t < \min\left\{\widetilde{\tau}_{[i]}, \widetilde{\tau}_{[i-1]}^{\text{good}}\right\}\right\},$$

if such a time exists, and $\widetilde{\tau}^{\text{good}}_{[i]} = -\infty$ otherwise. Let i^* be the index of the largest time $\widetilde{\tau}_{[i^*]}$ such that $\widetilde{\tau}^{\text{good}}_{[i^*]}$ does not exist, i.e.,

$$i^* = \min\left\{i \in [|S(\tau)^c|] : S(\tau) \cap \left[1, \min\left\{\widetilde{\tau}_{[i]}, \widetilde{\tau}^{\text{good}}_{[i-1]}\right\}\right) = \emptyset\right\}.$$

Notice that $\widetilde{\tau}^{\text{good}}_{[i]} = -\infty$ for every $i \geq i^*$, and $\widetilde{\tau}^{\text{good}}_{[i]} \in S(\tau)$ otherwise. Notice that, for every $i < i^*$, we have that

$$\widetilde{\tau}_{[i]} - \widetilde{\tau}^{\text{good}}_{[i]} \leq |S(\tau)^c|. \tag{21}$$

Indeed, this follows by first decomposing

$$\widetilde{\tau}_{[i]} - \widetilde{\tau}^{\text{good}}_{[i]} = |(\widetilde{\tau}^{\text{good}}_{[i]}, \widetilde{\tau}_{[i]}) \cap S(\tau)^c| + |(\widetilde{\tau}^{\text{good}}_{[i]}, \widetilde{\tau}_{[i]}) \cap S(\tau)| + 1.$$

Notice that $|(\widetilde{\tau}^{\text{good}}_{[i]}, \widetilde{\tau}_{[i]}) \cap S(\tau)| \leq i - 1$, since there are exactly $i - 1$ times $\widetilde{\tau}_{[i']} > \widetilde{\tau}_{[i]}$, and each has a time $\widetilde{\tau}^{\text{good}}_{[i']} \in S(\tau)$, which may lie on that interval. Note that there cannot be more than $i - 1$ times $t \in S(\tau)$ on this interval, since this would violate our choice of $\widetilde{\tau}^{\text{good}}_{[i]}$ as the largest time in $S(\tau)$ smaller than $\widetilde{\tau}_{[i]}$ which wasn't assigned to an earlier $\widetilde{\tau}_{[i']}$. Further, notice that $|(\widetilde{\tau}^{\text{good}}_{[i]}, \widetilde{\tau}_{[i]}) \cap S(\tau)^c| \leq |S(\tau)^c| - i$ by definition of $\widetilde{\tau}_{[i]}$. Combining these two bounds yields the claim.

Next, notice that, for every $i \geq i^*$,

$$\widetilde{\tau}_{[i]} \leq \widetilde{\tau}_{[i^*]} \leq |S(\tau)^c|, \tag{22}$$

where the first inequality is by definition of $\widetilde{\tau}_{[i]}$. To see the second inequality, we follow a similar argument as before. Indeed, observe that

$$\widetilde{\tau}_{[i^*]} = |[1, \widetilde{\tau}_{[i^*]}) \cap S(\tau)^c| + |[1, \widetilde{\tau}_{[i^*]}) \cap S(\tau)| + 1.$$

By definition of i^*, $|[1, \widetilde{\tau}_{[i^*]}) \cap S(\tau)| \leq i^* - 1$, since the only times $t \in S(\tau)$ on this interval can be $\widetilde{\tau}^{\text{good}}_{[1]}, \ldots, \widetilde{\tau}^{\text{good}}_{[i^*-1]}$ by definition of i^* (otherwise, we would have $\widetilde{\tau}^{\text{good}}_{[i^*]} > -\infty$). Further, $|[1, \widetilde{\tau}_{[i^*]}) \cap S(\tau)^c| \leq |S(\tau)^c| - i^*$ by definition of $\widetilde{\tau}_{[i^*]}$. Combining these two bounds yields the claim.

As a result, we have the following:

$$\begin{aligned}
&\sum_{t \in S(\tau)^c} \|\nabla F(\mathbf{w}_t)\|^2 = \sum_{i=1}^{i^*-1} \left\|\nabla F(\mathbf{w}_{\widetilde{\tau}_{[i]}})\right\|^2 + \sum_{i=i^*}^{|S(\tau)^c|} \left\|\nabla F(\mathbf{w}_{\widetilde{\tau}_{[i]}})\right\|^2 \\
&\leq \sum_{i=1}^{i^*-1} 2c_k^2 \left\|\nabla F(\mathbf{w}_{\widetilde{\tau}^{\text{good}}_{[i]}})\right\|^2 + 2c_k'^2 \left\|\mathbf{w}_{\widetilde{\tau}_{[i]}} - \mathbf{w}_{\widetilde{\tau}^{\text{good}}_{[i]}}\right\|^{2(k-1)} \\
&\quad + \sum_{i=i^*}^{|S(\tau)^c|} 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 + 2\max\left\{c_k'^2 \left\|\mathbf{w}_{\widetilde{\tau}_{[i]}} - \mathbf{w}_1\right\|^{2(k-1)}, L_0^2 \left\|\mathbf{w}_{\widetilde{\tau}_{[i]}} - \mathbf{w}_1\right\|^2\right\}.
\end{aligned}$$

Hence, by Fact 16, we have the bound

$$\begin{aligned}
&\sum_{t\in S(\tau)^c} \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \sum_{t\in S(\tau)} 2c_k^2 \|\nabla F(\mathbf{w}_t)\|^2 + \sum_{i=1}^{i^*-1} 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}(\widetilde{\tau}_{[i]} - \widetilde{\tau}_{[i]}^{\text{good}})^{2(k-1)} \\
&\quad + 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 |S(\tau)^c| + \sum_{i=i^*}^{|S(\tau)^c|} 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}\widetilde{\tau}_{[i^*]}^{2(k-1)} \\
&\leq \sum_{t\in S(\tau)} 2c_k^2 \|\nabla F(\mathbf{w}_t)\|^2 + 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}(i^*-1)|S(\tau)^c|^{2(k-1)} \\
&\quad + 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 |S(\tau)^c| + 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}(|S(\tau)^c| - (i^*-1))|S(\tau)^c|^{2(k-1)} \\
&= \sum_{t\in S(\tau)} 2c_k^2 \|\nabla F(\mathbf{w}_t)\|^2 + 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}|S(\tau)^c|^{2k-1} + 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 |S(\tau)^c|,
\end{aligned}$$

which is the first stated bound.

To obtain the second, we apply the first, where we choose $\tau := \tau_{T+1}(\delta)$ (the stopping time from Definition 9) and $S(\tau) := S_{\text{good}}(\tau_{T+1}(\delta))$ (the set of "good" times before $\tau_{T+1}(\delta)$ from Definition 21). Thus, for any $\widetilde{S}(\tau_{T+1}(\delta)) \subseteq S_{\text{good}}(\tau_{T+1}(\delta))$ for which $\mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|\right] \leq (1+n_{\text{comp}})\mathbb{E}\left[|S_{\text{good}}(\tau_{T+1}(\delta))|\right]$, we conclude that:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))^c} \|\nabla F(\mathbf{w}_t)\|^2\right] \\
&\leq 2c_k^2\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2\right] \\
&\quad + 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}(n_{\text{comp}}+1)^{2k-1}\mathbb{E}\left[|\widetilde{S}(\tau_{T+1}(\delta))^c|^{2k-1}\right] \\
&\quad + 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 (n_{\text{comp}}+1)\mathbb{E}\left[|S_{\text{good}}(\tau_{T+1}(\delta))^c|\right] \\
&\leq 2c_k^2\mathbb{E}\left[\sum_{t\in\widetilde{S}(\tau_{T+1}(\delta))} \|\nabla F(\mathbf{w}_t)\|^2\right] \\
&\quad + 2\max\left\{c_k'^2\eta^{2(k-1)}, L_0^2\eta^2\right\}(n_{\text{comp}}+1)^{2k-1}\left(\frac{2k\sigma_1^2\log(f(\tau_T(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^{2k-1} \\
&\quad + 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 (n_{\text{comp}}+1)\left(\frac{2\sigma_1^2\log(f(\tau_T(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right),
\end{aligned}$$

where in the second inequality, we applied Lemma 12. Thus, we obtain the claimed result. ■

C.4. Applying Theorem 26 to polynomially-bounded functions with no restriction on σ_1

Now that we have shown in the previous results how to upper bound $\mathrm{comp}(\tau_{T+1}(\delta))$, the sum of "bad" gradients, and the moments of the size of the "bad" set, we are now ready to establish our second main result: a convergence guarantee for functions satisfying (L_0, L_1)-smoothness and Definition 4, which holds for arbitrary $\sigma_0, \sigma_1 \geq 0$.

Corollary 32 (of Theorem 26; Formal statement of Theorem 5) *Fix any $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''' \in (0,1)$ satisfying $\varepsilon + \varepsilon' + \varepsilon'' + \varepsilon''' < 1$. Consider (AG-Norm) with any parameters $\eta \leq {}^{2\varepsilon'}/_{L_1(4+\sigma_1^2)}$ and $b_0^2 > 0$, running for $T \geq 1$ time steps on an objective function satisfying Assumption 2 as well as Definition 4 for some constants $k \geq 2, c_k \geq 1, c_k' > 0$. Suppose that the stochastic gradient oracle satisfies Assumption 4 for any $\sigma_0, \sigma_1 \geq 0$. Then, for any $\delta' \in (0,1)$ and $T \geq 1$, with probability at least $1 - \delta' - \frac{4(1+n_{\mathrm{comp}})\sigma_1^2 \log(f(T))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2 T}$, (AG-Norm) satisfies:*

$$\begin{aligned}
&\min_{t\in[T]} \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \frac{32(1+c_{\mathrm{B2}})b(T)^2}{\eta^2(\delta')^2 T} + \frac{16b(T)}{\eta(\delta')^2 T}\sqrt{b_0^2 + \sigma_0^2 + 2(1+\sigma_1^2)c_{\mathrm{B1}} + \frac{4b(T)}{\eta}\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{b_0^2 + 2\eta^2 L_0^2}} \\
&\quad + \frac{32 b(T)^{3/2}}{\eta^{3/2}(\delta')^{2.25} T^{3/4}}\sqrt{2\sigma_1^2(1+c_{\mathrm{B2}})\sqrt{(1+c_{\mathrm{L}}+\sigma_1^2)c_{\mathrm{B1}}}} \\
&\quad + \frac{16b(T)}{\eta(\delta')^2\sqrt{T}}\sqrt{2\sigma_0^2 + \frac{8\sigma_1^2(1+c_{\mathrm{B2}})b(T)}{\eta\sqrt{\delta'}}\left(\frac{2(1+c_{\mathrm{B2}})(1+c_{\mathrm{L}}+\sigma_1^2)b(T)}{\eta\sqrt{\delta'}} + \sigma_0\right)},
\end{aligned}$$

where $n_{\mathrm{comp}} = \left\lceil \frac{4c_k^3(\sigma_1 - (1-\varepsilon-\varepsilon'))_+}{\varepsilon'''} \right\rceil$, $c_{\mathrm{L}} = 2(1+\eta L_1)^2$, $c_{\mathrm{B2}} = 2c_k^2$,

$$\begin{aligned}
b(T) &:= \frac{1}{\varepsilon'''}\left(F(\mathbf{w}_1) - F^* + 2\widetilde{c}_0 \log\left(\frac{(2+\sigma_1^2)\widetilde{c}_0 T}{\eta\varepsilon'' b_0}\right) + \frac{2\eta\varepsilon''\sigma_0}{(2+\sigma_1^2)} + \mathrm{comp}(T)\right) \\
\mathrm{comp}(T) &= \eta(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ c_k \|\nabla F(\mathbf{w}_1)\| \ell_1(T) \\
&\quad + \eta n_{\mathrm{comp}}^{k-1} \max\left\{c_k'\eta^{k-1}, L_0\eta\right\}((2c_k+1)\sigma_1 + 1/2)\,\ell_k(T) \\
c_{\mathrm{B1}} &= 2c_k^2 \|\nabla F(\mathbf{w}_1)\|^2 (n_{\mathrm{comp}}+1)\ell_1(T) + 2\max\left\{c_k'\eta^{k-1}, L_0\eta\right\}^2 (n_{\mathrm{comp}}+1)^{2k-1}\ell_{2k-1}(T), \\
\ell_k(T) &= \left(\frac{(k+1)\sigma_1^2 \log(f(T))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^k \\
f(T) &= e + 4eT\frac{\sigma_0^2 T + (1+\sigma_1^2+c_{\mathrm{L}})\left(2Tc_k^2\|\nabla F(\mathbf{w}_1)\|^2 + 2\max\left\{c_k'\eta^{k-1}, L_0\eta\right\}^2 T^{2(k-1)}\right)}{b_0^2\delta'},
\end{aligned}$$

and $\widetilde{c}_0 = \frac{\eta\sigma_0}{2\varepsilon} + \eta^2\frac{L_0+\sigma_0 L_1}{2}$ *(where we use the notation* $(x)_+ := \max\{0, x\}$*).*

Proof We apply Theorem 26 as follows. First, we observe that, as a consequence of Lemma 11, together with the bound on $\mathbb{E}\left[|S_{\text{good}}(\tau_{T+1}(\delta))^c|^k\right]$ from Lemma 12, we have that:

$$\begin{aligned}
&\text{comp}(\tau_{T+1}(\delta)) \\
&\leq \eta(\sigma_1 - (1-\varepsilon-\varepsilon'))_+ c_k \left\|\nabla F(\mathbf{w}_1)\right\| \left(\frac{2\sigma_1^2 \log(f(\tau_T(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right) \\
&\quad + \eta n_{\text{comp}}^{k-1} \max\left\{c_k' \eta^{k-1}, L_0 \eta\right\} \left(\frac{(k+1)\sigma_1^2 \log(f(\tau_T(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^k \left((\sigma_1 - (1-\varepsilon-\varepsilon'))_+ + \frac{\varepsilon''' n_{\text{comp}}}{2c_k^3}\right).
\end{aligned}$$

Next, by Lemma 31, we know that

$$\sum_{t \in \widetilde{S}(\tau_{T+1}(\delta))^c} \left\|\nabla F(\mathbf{w}_t)\right\|^2 \leq \mathrm{B}_1 + c_{\mathrm{B}2} \sum_{t \in \widetilde{S}(\tau_{T+1}(\delta))} \left\|\nabla F(\mathbf{w}_t)\right\|^2,$$

where

$$\begin{aligned}
\mathbb{E}\left[\mathrm{B}_1\right] \leq c_{\mathrm{B}1} &= 2\max\left\{c_k'^2 \eta^{2(k-1)}, L_0^2 \eta^2\right\} (n_{\text{comp}}+1)^{2k-1} \left(\frac{2k\sigma_1^2 \log(f(\tau_T(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right)^{2k-1} \\
&\quad + 2c_k^2 \left\|\nabla F(\mathbf{w}_1)\right\|^2 (n_{\text{comp}}+1) \left(\frac{2\sigma_1^2 \log(f(\tau_T(\delta)))}{(1-(\varepsilon+\varepsilon'+\varepsilon''+\varepsilon'''))^2}\right), \\
c_{\mathrm{B}2} &= 2c_k^2.
\end{aligned}$$

Thus, the conditions to apply Theorem 26 are satisfied, and we obtain the convergence rate. ■

Appendix D. Many common algorithms for (L_0, L_1)-smooth optimization can diverge in the presence of multiplicative noise

In this section, we consider the convergence behavior of several natural candidate algorithms which have been studied in the literature on (L_0, L_1)-smooth optimization. These algorithms take the form $\mathbf{w}_{t+1} = \mathbf{w}_t - \boldsymbol{u}_t$, where $\boldsymbol{u}_t$ takes a number of different forms, including: in Normalized SGD:

$$\boldsymbol{u}_t = \eta \frac{\boldsymbol{g}_t}{\gamma + \|\boldsymbol{g}_t\|}, \tag{NormSGD}$$

Clipped SGD:

$$\boldsymbol{u}_t = \eta \frac{\boldsymbol{g}_t}{\max\{\gamma, \|\boldsymbol{g}_t\|\}} \tag{ClippedSGD}$$

and Sign-SGD with Momentum (operations performed element-wise):

$$\boldsymbol{u}_t = \eta \frac{\boldsymbol{m}_t}{|\boldsymbol{m}_t|} \quad \text{where} \quad \boldsymbol{m}_0 = \mathbf{0}, \quad \boldsymbol{m}_t = \beta \boldsymbol{m}_{t-1} + (1-\beta)\boldsymbol{g}_t \tag{SignSGD-M}$$

Zhang et al. (2020b,a); Crawshaw et al. (2022) prove $\mathcal{O}(1/\sqrt{T})$ convergence of these algorithms in the setting of (Bounded-supp). In this section, we show that these step-size choices for (L_0, L_1)-smooth optimization fail under (Affine-var), despite working in the noiseless and (Bounded-supp) settings. Our negative results rely on the following stochastic gradient oracle construction:

Proposition 13 (A stochastic gradient oracle satisfying Assumption 4) *Fix any $\sigma_0, \sigma_1 \geq 0$, and consider the following stochastic gradient oracle: fix any $\varepsilon \geq 0$, and let, for every $\mathbf{w} \in \mathbb{R}^d$:*

$$\xi_{mult}(\mathbf{w}) = \begin{cases} \left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) & \text{w.p. } \delta = \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2} \\ -\varepsilon & \text{w.p. } 1-\delta \end{cases} \quad \text{and} \quad \xi_{add}(\mathbf{w}) \sim \mathcal{N}(0, \sigma_0^2 I_{d\times d}).$$

We can then take the output of the oracle to be $\boldsymbol{g}(\mathbf{w}) := \xi_{add}(\mathbf{w}) + \xi_{mult}(\mathbf{w})\nabla F(\mathbf{w})$. Then, this construction satisfies Assumptions 3 and 4 with the specified σ_0 and σ_1.

Proof Fix any $\varepsilon, \sigma_1 \geq 0$. We begin by establishing that Assumption 3 holds for our construction of $\boldsymbol{g}(\mathbf{w})$. Begin by denoting $\delta = {(1+\varepsilon)^2}/{((1+\varepsilon)^2+\sigma_1^2)}$. Under this notation, we have that

$$1 + \frac{\sigma_1^2}{1+\varepsilon} = \frac{(1+\varepsilon)^2 + \sigma_1^2(1+\varepsilon)}{(1+\varepsilon)^2} = \frac{1}{\delta} + \varepsilon\frac{(1-\delta)}{\delta} = \frac{1+\varepsilon(1-\delta)}{\delta}.$$

Therefore, it follows that

$$\mathbb{E}\left[\xi_{\text{mult}}(\mathbf{w})\right] = (-\varepsilon(1-\delta) + \varepsilon(1-\delta)) = 1.$$

Further, $\mathbb{E}\left[\xi_{\text{add}}(\mathbf{w})\right] = 0$ by construction. Therefore, $\mathbb{E}\left[\boldsymbol{g}(\mathbf{w})\right] = \nabla F(\mathbf{w})$, which establishes Assumption 3. As for Assumption 4, denote $c = 1 + \varepsilon$, then we have that

$$\begin{aligned} \mathbb{E}\left[\xi_{\text{mult}}(\mathbf{w})^2\right] &= (c-1)^2\frac{\sigma_1^2}{c^2+\sigma_1^2} + \left(1 + \frac{\sigma_1^2}{c}\right)^2\frac{c^2}{c^2+\sigma_1^2} \\ &= \frac{(c^2+1-2c)\sigma_1^2 + c^2 + \sigma_1^4 + 2c\sigma_1^2}{c^2+\sigma_1^2} \\ &= (1+\sigma_1^2). \end{aligned}$$

Further, $\mathbb{E}\left[\left\|\xi_{\text{add}}(\mathbf{w})\right\|^2\right] = \sigma_0^2$ by construction. Therefore, since $\xi_{\text{mult}}(\mathbf{w})$ and $\xi_{\text{add}}(\mathbf{w})$ are independent, we conclude that

$$\begin{aligned} \mathbb{E}\left[\left\|\boldsymbol{g}(\mathbf{w})\right\|^2\right] &= \mathbb{E}\left[\xi_{\text{mult}}(\mathbf{w})^2\right]\left\|\nabla F(\mathbf{w})\right\|^2 + \mathbb{E}\left[\left\|\xi_{\text{add}}(\mathbf{w})\right\|^2\right] + 2\mathbb{E}\left[\langle\xi_{\text{mult}}(\mathbf{w})\nabla F(\mathbf{w}), \xi_{\text{add}}(\mathbf{w})\rangle\right] \\ &= (1+\sigma_1^2)\left\|\nabla F(\mathbf{w})\right\|^2 + \sigma_0^2 + 2\left\langle\mathbb{E}\left[\xi_{\text{mult}}(\mathbf{w})\right]\nabla F(\mathbf{w}), \mathbb{E}\left[\xi_{\text{add}}(\mathbf{w})\right]\right\rangle \\ &= (1+\sigma_1^2)\left\|\nabla F(\mathbf{w})\right\|^2 + \sigma_0^2, \end{aligned}$$

which establishes Assumption 4 for any $\sigma_0, \sigma_1 \geq 0$. ■

D.1. Overview of main negative results

We establish all of the following negative results using the stochastic gradient oracle described in Proposition 13. Before stating our results, let us briefly discuss some intuition behind why one should expect (NormSGD), (ClippedSGD), and (SignSGD-M) to fail under Proposition 13. Consider the setting where $\sigma_1 \gg 1 + \varepsilon$. Then, notice that the stochastic gradient $\boldsymbol{g}_t$ only has the same sign as $\nabla F(\mathbf{w}_t)$ with roughly ${1}/{\sigma_1^2}$ probability. Otherwise, $\boldsymbol{g}_t$ has the opposite sign as $\nabla F(\mathbf{w}_t)$. Now, for an algorithm which incorporates the *magnitude* of the stochastic gradients together with the signs,

the oracle in Proposition 13 may not be so problematic – indeed, even though the updates with correct sign are somewhat "rare", they are also of significantly larger magnitude compared to the updates with proper sign. However, notice that (NormSGD), (ClippedSGD), and (SignSGD-M) are (effectively) unit step-length algorithms (at least, in the setting where $\|\boldsymbol{g}_t\| \geq \gamma$). Thus, in many parameter regimes, all of these algorithms effectively disregard the magnitude of the stochastic gradients and only use their signs. This results in a biased random walk which never finds an iterate better than the initial one with constant probability. We formalize this intuition in the following:

Lemma 33 (Informal statement of Lemma 35) *Fix any smoothness parameter $L_0 > 0$, initial gap $\Delta > 0$, and affine variance parameter $\sigma_1 > 2\sqrt{2}$. Suppose that either: (i) (SignSGD-M) is run with parameter $0 \leq \beta \leq 1 - 2\sqrt{2}/3 \approx 0.057$ and $\eta > 0$ for $T \geq 1$ time steps, or (ii) (NormSGD) or (ClippedSGD) is run with $0 \leq \gamma \leq \sqrt{\sigma_1^2 \Delta L_0}/2$ and $\eta > 0$ for $T \geq 1$ time steps, where, in either case, the algorithms are allowed an arbitrary initialization $x_1 \in \mathbb{R}$, and each of these parameters can depend on L_0, Δ and σ_1. Then, there exists a 1-dimensional $(L_0, 0)$-smooth function (which is also L_0-strongly convex) with $F(x_1) - \inf_{x \in \mathbb{R}} F(x) = \Delta$, and stochastic gradient oracle satisfying Assumptions 3 and 4 with $\sigma_0 = 0$ and the specified σ_1, and for which, with constant probability (independent of T), $\min_{t \in [T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2$.*

We note that the statement Lemma 33 follows from Lemma 35 by choosing the parameter $\varepsilon = \sigma_1/2\sqrt{2}$. The main takeaway here is that, for a reasonably wide range of parameters, (NormSGD), (ClippedSGD), and (SignSGD-M) can diverge in the affine variance setting, even for very simple smooth and strongly convex problems (in fact, even on a 1-dimensional quadratic function). In particular, this says that, whenever (NormSGD) is run with $\gamma = 0$ (or (SignSGD-M) with $\beta = 0$), then there is no parameter tuning with respect to η such that $\min_{t \in [T]} \|\nabla F(x_t)\|^2$ converges!

We also give a (weaker) negative result for the (AG-Norm) in the "large variance" regime. This result establishes that, whenever η is not carefully tuned with respect to both L_1 and σ_1, then the algorithm does not converge with constant probability. The intuition for this result is that, with constant probability, the first $\approx \sigma_1^2$ stochastic gradients all have the wrong sign. Whenever σ_1 is "large" (i.e., scaling as $\operatorname{poly}\log(T)$), then after only $\operatorname{poly}\log(T)$ steps, the algorithm can reach an objective value which is $\operatorname{poly}(T)$-times larger than the initial condition. Further, after reaching such a large gradient value, the step sizes are always too small for the algorithm to recover from these wrong initial steps. This is because the (AG-Norm) updates are normalized by the large previous gradients.

Lemma 34 (Informal statement of Lemma 39) *Fix any $L_1 > 0$, time horizon $T > 1$, and affine variance parameter*

$$\sigma_1 \geq \max\left\{ \left(\frac{4(1+\sqrt{2})^2 - 2}{\log(4/3)} \right)^{2/3}, \left(\frac{16 \log(T-1))^2}{\log(4/3)} \right)^2 \right\}.$$

Suppose that (AG-Norm) is initialized at $x_1 \in \mathbb{R}$ and run with any parameters $\eta \geq 1/(2L_1\sqrt{\sigma_1})$ and $0 < b_0^2 \leq \sqrt{\sigma_1} L_1^2 \exp(2L_1 x_1)$ (where these parameter choices may depend on L_1). Then, there exists a 1-dimensional $(0, (e-1)L_1)$-smooth function such that $\inf_{x \in \mathbb{R}} F(x) = 0$, and a stochastic gradient oracle satisfying Assumptions 3 and 4 with $\sigma_0 = 0$ and the specified σ_1, for which, with probability at least $3/4$, $\min_{t \in [T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2$.

We note that the statement Lemma 34 follows from Lemma 39 by choosing the parameters $\alpha = 1/(2\sqrt{\sigma_1})$, $\varepsilon = \sqrt[4]{\sigma_1} - 1$, and $\delta = 1/4$. Let us compare the negative result in Lemma 34 with the convergence result in the L_0-smooth regime for the same algorithm from (Faw et al., 2022). Indeed, their main result was that a $\widetilde{\mathcal{O}}(1/\sqrt{T})$ convergence rate is achievable without tuning the parameters of the algorithm with respect to σ_0, σ_1, or L_0. Since their convergence rate depends only polynomially on σ_1, this rate is maintained (up to poly-logarithmic factors) even when $\sigma_1 = \operatorname{poly}\log(T)$, without adjusting the parameters η or b_0 of the algorithm. By contrast, Lemma 34 tells us that, in the (L_0, L_1)-smooth regime, such a result is no longer possible. Indeed, if η is not sufficiently small, then the algorithm does not converge with constant probability when $\sigma_1^2 \gtrsim \operatorname{poly}\log(T)$!

D.2. Full statement and proof of negative results for (SignSGD-M), (NormSGD), and (ClippedSGD)

Here, we give the complete negative result for (SignSGD-M), (NormSGD), and (ClippedSGD), and formalize the intuition given there.

Lemma 35 (Formal statement of Lemma 33) *Fix any $L_0 > 0$, $\varepsilon > 0$, $\sigma_1^2 > (1+\varepsilon)^2$, and $\Delta > 0$. Let $x_1 \in \mathbb{R}$, $\eta > 0$, $\gamma \in [0, \varepsilon\sqrt{2\Delta L_0}]$, $\beta \in \left[0, 1 - \sqrt{1 - \frac{\varepsilon}{1+\varepsilon+\sigma_1^2/(1+\varepsilon)}}\right) \supset \left[0, \frac{\varepsilon}{2(1+\varepsilon+\sigma_1^2/(1+\varepsilon))}\right)$, and $T \geq 1$ be arbitrary parameters (possibly dependent on L_0, ε, σ_1, and Δ). For any $t \in [T]$, consider the (one-dimensional) process $\{x_t\}_{t\geq 1}$ given in (SignSGD-M), (NormSGD), or (ClippedSGD). where, in the case that $m_t = 0$ (in the case of (SignSGD-M)) or $\mu + |g_t| = 0$ (in the case of (NormSGD)), $u_t \in \{\pm\eta\}$ may be chosen arbitrarily as a (possibly randomized) function of $\{g_1, \ldots, g_t\}$. Then, assuming that $\sigma_1^2 > (1 + \gamma/\varepsilon\sqrt{2\Delta L_0})(1+\varepsilon)^2 \in [(1+\varepsilon)^2, 2(1+\varepsilon)^2]$, there exists an 1-dimensional $(L_0, 0)$-smooth function (which is also L_0-strongly convex) with $F(x_1) - \inf_{x\in\mathbb{R}} F(x) = \Delta$, and stochastic gradient oracle which outputs stochastic gradients g_t of $\nabla F(x_t)$ which satisfy Assumptions 3 and 4, and such that:*

$$\Pr\left[\min_{t\in[T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2\right] \geq (1-\delta)^{t_0+1},$$

where

$$t_0 = \left\lceil \frac{\sqrt{2\delta(1-\delta)\log(1+2/\delta)}}{\delta_0}\left(1 + \frac{2\delta(1-\delta)}{(\delta-\delta_0)^2}\log\left(\frac{4(1-\delta)}{(\delta_0-\delta)^2}\right)\right)\right\rceil,$$

and $\delta = \frac{1}{(1+\sigma_1^2/(1+\varepsilon)^2)} < \frac{1}{1+1/\lambda_{\text{clip}}} = \delta_0 \in [1/3, 1/2]$ and $1/\lambda_{\text{clip}} = 1 + \gamma/(\varepsilon\sqrt{2\Delta L_0}) \in [1, 2]$.

Proof Let us choose, for arbitrary $L_0 > 0$ and $\Delta > 0$, the $(L_0, 0)$-smooth objective $F(x) = L_0/2\, x^2$, and assume without loss of generality that $x_1 = -\sqrt{2\Delta/L_0}$ (indeed, if this is not the case, then we can always translate the function $F(x)$ to be $F(x) = L_0/2(x - x_1 - \sqrt{2\Delta/L_0})^2$, and our arguments remain unchanged). Notice that $F(x_1) - F^* = F(x_1) = \Delta$.

Consider, for any $\varepsilon > 0$ and $\sigma_1^2 > (1+\varepsilon)^2$, the stochastic gradient oracle from Proposition 13, i.e.,

$$g(x) := \begin{cases} \left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) L_0 x & \text{w.p. } \frac{1}{1+\frac{\sigma_1^2}{(1+\varepsilon)^2}} := \delta \\ -\varepsilon L_0 x & \text{w.p. } 1 - \frac{1}{1+\frac{\sigma_1^2}{(1+\varepsilon)^2}} = 1 - \delta, \end{cases}$$

where the multiplicative noise is sampled i.i.d for each x. Since $\nabla F(x) = L_0 x$, this construction satisfies Assumptions 3 and 4 by Proposition 13. Further, denoting $x_{\text{clip}} := -\gamma/\varepsilon L_0$, our assumption that $\gamma \leq \varepsilon\sqrt{2\Delta L_0} = -\varepsilon L_0 x_1$ and $\sigma_1^2 > (1+\varepsilon)^2$ (and thus also $\varepsilon < 1 + \sigma_1^2/(1+\varepsilon)$) ensures:

$$x_1 \leq x_{\text{clip}} < 0 \quad \text{and} \quad |g(x_{\text{clip}})| \geq \varepsilon L_0 \, |x_{\text{clip}}| = \gamma. \tag{23}$$

Let τ^* be the first time when an iterate becomes larger than the original one, i.e.,

$$\tau^* = \min\left\{t > 1 : x_1 \leq x_t\right\}.$$

Notice that this implies that, for any $1 \leq t < \tau^*$:

$$x_t \leq x_1 \leq x_{\text{clip}} < 0 \quad \text{and} \quad |g_t| \geq \gamma \quad \text{and} \quad \|\nabla F(x_t)\|^2 \geq \|\nabla F(x_1)\|^2. \tag{24}$$

This guarantees that, before τ^*, (i) the iterates are always to the left of the minimizer, (ii) that the algorithm (ClippedSGD) never "clips" (i.e., $u_t = \eta g_t/|g_t|$), and (iii), $\min_{t<\tau^*} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2$ (i.e., the algorithm never achieves any nontrivial target minimization criterion). Additionally, it must be the case that:

$$u_{\tau^*-1} < 0, \tag{25}$$

since $x_{\tau^*-1} < x_1$ and $x_{\tau^*} = x_{\tau^*-1} - u_{\tau^*-1} \geq x_1$.

Now, let us distinguish the updates of (SignSGD-M), (NormSGD), and (ClippedSGD) as $(x_t(1), u_t(1))$, $(x_t(2), u_t(2))$, and $(x_t(3), u_t(3))$, respectively. Now, instead of reasoning about the dynamics of each of these algorithms individually, we instead reason about an algorithm with simpler dynamics, and draw conclusions about each of these processes via a stochastic dominance argument.

To do this, we utilize the coupling of these algorithms defined in Lemma 37 – namely, we let $x_1(i) = x_1 = \sqrt{2\Delta/L_0}$ (as discussed above), and $g(x_t(i)) = \xi_{\text{mult},t} \nabla F(x_t(i))$ for every i, where $\xi_{\text{mult},t}$ is $-\varepsilon$ with probability $1-\delta$, and $1 + \sigma_1^2/(1+\varepsilon)$ otherwise. That is, each process starts from the same initial iterate, and receives the same multiplicative noise on the stochastic gradient at time t.

Similarly, let us define the "simpler" comparison process as:

$$u_s(4) = \begin{cases} \lambda_{\text{clip}}\eta & \text{if } \xi_{\text{mult},s} = -\varepsilon \\ -\eta & \text{o.w.} \end{cases} \quad \text{and} \quad \lambda_{\text{clip}} := \frac{1}{1 + \frac{\gamma}{\varepsilon L_0 |x_1|}} = \frac{1}{1 + \frac{\gamma}{\varepsilon\sqrt{2\Delta L_0}}} \in [1/2, 1]$$

and take $x_1(4) = x_1$ and $x_{t+1}(4) = x_t(4) - u_t(4)$. Now, denote $\tau^*(i)$ as the stopping time from (23) corresponding to the process $i \in [4]$. Then, by Lemma 37, we have that, under our coupling of these algorithms, $\tau^*(4) \leq \min_{i\in[3]} \tau^*(i)$, which implies that, for each algorithm $i \in [3]$:

$$\Pr\left[\min_{t\in[T]} \|\nabla F(x_t(i))\|^2 = \|\nabla F(x_1)\|^2\right] \geq \Pr\left[\tau^*(i) > T\right] \geq \Pr\left[\tau^*(4) > T\right],$$

where the first inequality follows from (24). Thus, to lower bound the failure probability of algorithm i, it suffices to lower bound $\Pr[\tau^*(4) > T]$, and thus to reason only about the dynamics of this "simpler" process.

By Lemma 38, we have that, for any $t_0 \geq 0$:

$$\Pr\left[\tau^*(4) > T\right] \geq (1-\delta)^{t_0}\left(1 - \sum_{t=t_0+2}^{T} \Pr\left[\frac{1}{t-t_0-1}(X_t - \mathbb{E}\left[X_t\right]) \leq -\left(\delta_0 - \delta\right) - \frac{\delta_0 t_0}{t-t_0-1}\right]\right),$$

where X_t is a sum of $t - t_0 - 1$ i.i.d Bernoulli random variables, each with mean $1 - \delta = 1 - \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2} > 1/2$ (since, by assumption, $\sigma_1 > (1+\varepsilon)$), and $\delta_0 = \lambda_{\text{clip}}/(1+\lambda_{\text{clip}})$. We may therefore apply the Chernoff-Hoeffding inequality (Hoeffding, 1963, Theorem 1, Eq. (2.2)) to obtain:

$$\begin{aligned}
&\Pr\left[\frac{1}{t-t_0-1}(X_t - \mathbb{E}\left[X_t\right]) \leq -\left(\delta_0 - \delta\right) - \frac{\lambda_{\text{clip}} t_0}{(1+\lambda_{\text{clip}})(t-t_0-1)}\right] \\
&\leq \exp\left(-\frac{t-t_0-1}{2\delta(1-\delta)}\left(\delta_0 - \delta + \frac{\delta_0 t_0}{t-t_0-1}\right)^2\right).
\end{aligned}$$

Notice that, since $\sigma_1^2 > {(1+\varepsilon)^2}/{\lambda_{\text{clip}}} = (1+\varepsilon)^2(1+{\gamma}/{\varepsilon\sqrt{2\Delta L_0}})$, $\delta_0 - \delta = \frac{1}{2+\frac{\gamma}{\varepsilon\sqrt{2\Delta L_0}}} - \frac{1}{1+\frac{\sigma_1^2}{(1+\varepsilon)^2}} > 0$, which implies the above bound is always nontrivial. Thus, we can use that bound to obtain, for any $\ell > 0$:

$$\begin{aligned}
&\sum_{t=t_0+2}^{T} \Pr\left[\frac{1}{t-t_0-1}(X_t - \mathbb{E}\left[X_t\right]) \leq -\left(\delta_0 - \delta\right) - \frac{\delta_0 t_0}{t-t_0-1}\right] \\
&\leq \sum_{t=t_0+2}^{t_0+1+\left\lfloor\frac{\delta_0 t_0}{\ell}\right\rfloor} \exp\left(-\frac{t-t_0-1}{2\delta(1-\delta)}\left(\delta_0 - \delta + \ell\right)^2\right) + \sum_{t=t_0+2+\left\lfloor\frac{\delta_0 t_0}{\ell}\right\rfloor}^{T} \exp\left(-\frac{t-t_0-1}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right) \\
&= \sum_{t=0}^{\left\lfloor\frac{\delta_0 t_0}{\ell}\right\rfloor - 1} \exp\left(-\frac{t+1}{2\delta(1-\delta)}\left(\delta_0 - \delta + \ell\right)^2\right) + \sum_{t=0}^{T-t_0-2-\left\lfloor\frac{\delta_0 t_0}{\ell}\right\rfloor} \exp\left(-\frac{t+1+\left\lfloor\frac{\delta_0 t_0}{\ell}\right\rfloor}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right)
\end{aligned}$$

Thus, using the geometric summation formula, we can bound the above summations as:

$$\begin{aligned}
&\sum_{t=t_0+2}^{T} \Pr\left[\frac{1}{t-t_0-1}(X_t - \mathbb{E}\left[X_t\right]) \leq -\left(\delta_0 - \delta\right) - \frac{\delta_0 t_0}{t-t_0-1}\right] \\
&\leq \exp\left(-\frac{1}{2\delta(1-\delta)}\left(\delta_0 - \delta + \ell\right)^2\right) \frac{1 - \exp\left(-\frac{\left\lfloor \frac{\delta_0 t_0}{\ell} \right\rfloor}{2\delta(1-\delta)}\left(\delta_0 - \delta + \ell\right)^2\right)}{1 - \exp\left(-\frac{1}{2\delta(1-\delta)}\left(\delta_0 - \delta + \ell\right)^2\right)} \\
&\quad + \exp\left(-\frac{1 + \left\lfloor \frac{\delta_0 t_0}{\ell} \right\rfloor}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right) \frac{1}{1 - \exp\left(-\frac{1}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right)} \\
&\leq \frac{\exp\left(-\frac{\ell^2}{2\delta(1-\delta)}\right)}{1 - \exp\left(-\frac{\ell^2}{2\delta(1-\delta)}\right)} \left(1 - \exp\left(-\frac{\left\lfloor \frac{\delta_0 t_0}{\ell} \right\rfloor}{2\delta(1-\delta)}\left(\delta_0 - \delta + \ell\right)^2\right)\right) \\
&\quad + \frac{\exp\left(-\frac{1}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right)}{1 - \exp\left(-\frac{1}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right)} \exp\left(-\frac{\left\lfloor \frac{\delta_0 t_0}{\ell} \right\rfloor}{2\delta(1-\delta)}\left(\delta_0 - \delta\right)^2\right).
\end{aligned}$$

Now, let us focus on bounding the two terms in the above expression. To do this, we first observe that, for any $\mu, x > 0$ and $i \geq 0$,

$$\frac{\exp(-(1+i)x)}{1-\exp(-x)} \leq \mu \iff i \geq \frac{1}{x}\log\left(\frac{\exp(-x)}{\mu(1-\exp(-x))}\right) \quad \text{or} \quad i = 0 \quad \text{and} \quad x \geq \log\left(1 + \frac{1}{\mu}\right). \tag{26}$$

Taking $i = 0$ and $x = \ell^2/(2\delta(1-\delta))$, the above implies that the first term is upper-bounded by $\mu = \delta/2$ whenever $\ell \geq \sqrt{2\delta(1-\delta)\log(1+2/\delta)}$. For the second term, we take $i = \left\lfloor \frac{\delta_0 t_0}{\ell} \right\rfloor$, $x = (\delta_0-\delta)^2/(2\delta(1-\delta))$, and conclude that the second term is upper-bounded by $\mu = \delta/2$ whenever

$$\left\lfloor \frac{\delta_0 t_0}{\ell} \right\rfloor \geq \frac{2\delta(1-\delta)}{(\delta-\delta_0)^2}\log\left(\frac{2\exp\left(-\frac{(\delta_0-\delta)^2}{2\delta(1-\delta)}\right)}{\delta\left(1-\exp\left(-\frac{(\delta_0-\delta)^2}{2\delta(1-\delta)}\right)\right)}\right).$$

In particular, since $\exp(-x) < 1/(1+x)$ for any $x > 0$, and thus also $\exp(-x)/1-\exp(-x) < 1/x$, since $\lfloor x \rfloor > x - 1$, we have that the above inequality is satisfied whenever:

$$t_0 \geq \frac{\ell}{\delta_0}\left(1 + \frac{2\delta(1-\delta)}{(\delta-\delta_0)^2}\log\left(\frac{4(1-\delta)}{(\delta_0-\delta)^2}\right)\right).$$

Therefore, we can choose $\ell = \sqrt{2\delta(1-\delta)\log(1+2/\delta)}$ and:

$$t_0 = \left\lceil \frac{\sqrt{2\delta(1-\delta)\log(1+2/\delta)}}{\delta_0}\left(1 + \frac{2\delta(1-\delta)}{(\delta-\delta_0)^2}\log\left(\frac{4(1-\delta)}{(\delta_0-\delta)^2}\right)\right)\right\rceil,$$

and, combining our results, we conclude that, for any algorithm $i \in [3]$:

$$\Pr\left[\min_{t\in[T]} \|\nabla F(x_t(i))\|^2 = \|\nabla F(x_1)\|^2\right] > (1-\delta)^{t_0+1},$$

as claimed. ■

Lemma 36 *Consider the process $\{x_t\}_{t\geq 1}$ from (SignSGD-M) as defined in Lemma 35, where $x_1 < 0$, $F(x) := {}^{L_0}/_2\, x^2$ for some $L_0 > 0$, and g_t are the stochastic gradients output by the oracle from Proposition 13. Suppose that the parameter β of (SignSGD-M) satisfies:*

$$\beta \in \left[0, \frac{\varepsilon}{1+\varepsilon+\frac{\sigma_1^2}{1+\varepsilon}}\right).$$

Let $\tau^* = \min\{t > 1 : x_1 \leq x_t\}$. Then, if $t < \tau^*$ and $g_t = -\varepsilon\nabla F(x_t)$, then $u_t = \eta$.

Proof Recall that, by construction of the stochastic gradient oracle from Proposition 13, and since $\nabla F(x) = L_0 x$:

$$g(x) := \begin{cases} \left(1+\frac{\sigma_1^2}{1+\varepsilon}\right) L_0 x & \text{w.p. } \frac{1}{1+\frac{\sigma_1^2}{(1+\varepsilon)^2}} := \delta \\ -\varepsilon L_0 x & \text{w.p. } 1 - \frac{1}{1+\frac{\sigma_1^2}{(1+\varepsilon)^2}} = 1-\delta. \end{cases}$$

We wish to show that the process from (SignSGD-M) has the property that, whenever $g_t = -\varepsilon L_0 x_t$ and $x_s < 0$ for every $s \in [t]$, then $u_t = \eta$. We consider any initialization $x_1 < 0$, and denote τ^* to be the first time when an iterate becomes non-negative, i.e.,

$$\tau^* = \min\{t > 1 : x_1 \leq x_t\}.$$

Further, take:

$$\tau_0 := 0 \quad \text{and} \quad \tau_{i+1} := \min\{t > \tau_i : \boldsymbol{g}_t = -\varepsilon L_0 x_t \text{ or } u_t = \eta\}.$$

Notice that, since $u_t \in \{\pm\eta\}$ by definition of (SignSGD-M), and by construction of the stochastic gradient oracle:

$$\boldsymbol{g}_t = \left(1+\frac{\sigma_1^2}{1+\varepsilon}\right) L_0 x_t \quad \text{and} \quad u_t = -\eta \quad \forall t \in (\tau_i, \tau_{i+1})\ \forall i \geq 0. \tag{27}$$

Thus, it suffices to prove by induction that, for any $i \geq 0$, either $\tau_i \geq \tau^*$, or $u_{\tau_i} = \eta$, as long as

$$\beta < 1 - \sqrt{1 - \frac{\varepsilon}{1+\varepsilon+\frac{\sigma_1^2}{1+\varepsilon}}} = 1 - \sqrt{\frac{\varepsilon}{1+\varepsilon}}\delta.$$

For the base case of $i = 0$, we may assume without loss of generality that $u_{\tau_0} = u_0 = \eta$, since $m_0 = 0$ and the dynamics of the update rule do not depend on u_0 (i.e., the dynamics begin at time $t = 1$ and x_1 is the starting point of the process). Thus, the base case is true by construction.

Now, suppose the claim holds for some $i \geq 1$. Either $\tau_{i+1} \geq \tau^*$ or not. In the former case, the claim follows trivially, so let us assume that $\tau_{i+1} < \tau^*$. Since $\tau_i < \tau_{i+1} < \tau^*$ by construction, $u_{\tau_i} = \eta$ by the induction hypothesis. Further, let us assume that $g_{\tau_{i+1}} = -\varepsilon L_0 x_{\tau_{i+1}}$, since otherwise the claim again follows trivially by definition of τ_{i+1}. Thus, we can write:

$$\begin{aligned}
m_{\tau_{i+1}} &= \beta^{\tau_{i+1}-\tau_i} m_{\tau_i} + (1-\beta) \sum_{t=\tau_i+1}^{\tau_{i+1}} \beta^{\tau_{i+1}-t} g_t \\
&= \beta^{\tau_{i+1}-\tau_i} m_{\tau_i} + (1-\beta) L_0 \left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) \sum_{t=\tau_i+1}^{\tau_{i+1}-1} \beta^{\tau_{i+1}-t} (x_{\tau_i+1} + \eta(t - \tau_i - 1)) \\
&\quad - (1-\beta) L_0 \varepsilon (x_{\tau_i+1} + \eta(\tau_{i+1} - \tau_i - 1))
\end{aligned}$$

where the first equality is the definition of $m_{\tau_{i+1}}$. The second inequality follows from observation (27). Further, since $u_{\tau_i} = \eta$, then by definition of (SignSGD-M), either $m_{\tau_i} > 0$, or $m_{\tau_i} = 0$ and the algorithm chooses $u_{\tau_i} = \eta$. In either case, $m_{\tau_i} \geq 0$. Therefore, since, for $\beta \in [0, 1)$:

$$\begin{aligned}
&(1-\beta) \sum_{t=\tau_i+1}^{\tau_{i+1}-1} \beta^{\tau_{i+1}-t} (x_{\tau_i+1} + \eta(t - \tau_i - 1)) \\
&= \beta(x_{\tau_i+1} + \eta(\tau_{i+1} - \tau_i - 1)) - \beta^{\tau_{i+1}-\tau_i} x_{\tau_i+1} - \beta\eta \frac{1 - \beta^{\tau_{i+1}-\tau_i-1}}{1-\beta},
\end{aligned}$$

we obtain, using the fact that $x_{\tau_{i+1}} = x_{\tau_i+1} + \eta(\tau_{i+1} - (\tau_i + 1))$ and $m_{\tau_i} \geq 0$:

$$\begin{aligned}
\frac{m_{\tau_{i+1}}}{L_0} &= \frac{\beta^{\tau_{i+1}-\tau_i} m_{\tau_i}}{L_0} - \left((1-\beta)\varepsilon - \beta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\right)(x_{\tau_i+1} + \eta(\tau_{i+1} - \tau_i - 1)) \\
&\quad - \beta^{\tau_{i+1}-\tau_i}\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) x_{\tau_i+1} \\
&\quad - \beta\eta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) \frac{1 - \beta^{\tau_{i+1}-\tau_i-1}}{1-\beta} \\
&\geq -\left((1-\beta)\varepsilon - \beta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\right)(x_{\tau_{i+1}} + \eta) \\
&\quad - \beta^{\tau_{i+1}-\tau_i}\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right) x_{\tau_i+1} \\
&\quad + \eta\left((1-\beta)\varepsilon - \beta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\left(1 + \frac{1 - \beta^{\tau_{i+1}-\tau_i-1}}{1-\beta}\right)\right).
\end{aligned}$$

Thus, since $x_{\tau_i+1} \leq x_{\tau_{i+1}} < 0$, and since $\tau_{i+1} < \tau^*$ (which implies, since each update of (SignSGD-M) satisfies $u_t \in \{\pm\eta\}$ and by definition of τ^*, $x_{\tau_{i+1}} \leq x_{\tau^*-1} = x_1 - \eta < 0$), the above inequality implies that $m_{\tau_{i+1}} > 0$ as long as:

$$(1-\beta)\varepsilon - \beta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\left(1 + \frac{1 - \beta^{\tau_{i+1}-\tau_i-1}}{1-\beta}\right) > (1-\beta)\varepsilon - \beta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\left(1 + \frac{1}{1-\beta}\right) > 0.$$

Since we require $0 \leq \beta < 1$, the second inequality is equivalent to:

$$(1-\beta)^2\varepsilon > \left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\beta(2-\beta),$$

which is satisfied as long as:

$$\beta < 1 - \sqrt{\frac{1+\frac{\sigma_1^2}{1+\varepsilon}}{1+\varepsilon+\frac{\sigma_1^2}{1+\varepsilon}}} = 1 - \sqrt{1 - \frac{\varepsilon}{1+\varepsilon}\frac{1}{1+\frac{\sigma_1^2}{1+\varepsilon}}} = 1 - \sqrt{1-\frac{\varepsilon}{1+\varepsilon}\delta}.$$

Thus, since $\sqrt{1-x} < 1 - x/2$ for $0 < x \leq 1$, it suffices to choose β as:

$$\beta \leq \frac{\varepsilon}{2(1+\varepsilon)}\delta < 1 - \sqrt{1-\frac{\varepsilon}{1+\varepsilon}\delta}.$$

In this case, $m_{\tau_{i+1}} > 0$, and thus $u_t = \eta$, which establishes the induction step. Thus, for every $i \geq 0$, either $\tau_i \geq \tau^*$ or $u_{\tau_i} = \eta$, as claimed. ■

Lemma 37 *Let us recall the i.i.d random process $\{\xi_{mult,t}\}_{t\geq 1}$ from Proposition 13, where each $\xi_{mult,t}$ is $-\varepsilon$ with probability $1-\delta$, and $(1+\sigma_1^2/(1+\varepsilon))$ otherwise. Let us distinguish the three processes from Lemma 35 (Eqs. (SignSGD-M), (ClippedSGD) and (NormSGD)) as, respectively, $\{x_t(i)\}_{t\geq 1}$ for $i \in [3]$. Consider the coupling of these three processes, where $x_1(i) = x_1 := -\sqrt{2\Delta/L_0}$ for every $i \in [4]$, and for each $t \geq 1$ and $i \in [3]$, $g(x_t(i)) = \xi_{mult,t}\nabla F(x_t(i))$. Further, let us denote, for each $t \geq 1$:*

$$u_t(4) = \begin{cases} \lambda_{\text{clip}}\eta & \text{if } \xi_{mult,t} = -\varepsilon \\ -\eta & \text{o.w.} \end{cases} \quad \text{where} \quad \lambda_{\text{clip}} := \frac{1}{1+\frac{\gamma}{\varepsilon\sqrt{2\Delta L_0}}} \in [1/2, 1],$$

and take $x_1(4) = x_1$ and $x_{t+1}(4) = x_t(4) - u_t(4)$. Further, let, for each $i \in [4]$,

$$\tau^*(i) = \min\{t > 1 : x_1 \leq x_t(i)\}.$$

Then, under the constraints on parameters of the three algorithms as imposed in Lemma 35, we have that:

$$\tau^*(4) \leq \min_{i\in[3]} \tau^*(i).$$

Proof We claim that, for each $i \in [3]$, and any $t < \tau^*(i)$, $u_t(4) \leq u_t(i)$. Notice that, supposing this claim is true, then $\tau^*(4) \leq \tau^*(i)$ for each $i \in [3]$, since, by definition of $\tau^*(i)$:

$$x_1 \leq x_{\tau^*(i)}(i) = x_1 - \sum_{s=1}^{\tau^*(i)-1} u_s(i) \leq x_1 - \sum_{s=1}^{\tau^*(i)-1} u_s(4) = x_{\tau^*(i)}(4).$$

Thus, since $\tau^*(i)$ is the first time $t > 1$ for which $x_t(i) \geq x_1$, it follows that $\tau^*(4) \leq \tau^*(i)$. Having established this implication, it suffices to prove the claim for each of the $u_t(i)$s.

For the case of $i = 1$ (i.e., algorithm (SignSGD-M)), this follows immediately from Lemma 36, since this result tells us that whenever $t < \tau^*(1)$ and $\xi_{\text{mult},t} = -\varepsilon$, then $u_t(1) = \eta > \eta\lambda_{\text{clip}} = u_t(4)$. Otherwise, whenever $t < \tau^*(1)$ and $\xi_{\text{mult},t} = (1 + \sigma_1^2/(1+\varepsilon))$, then by construction, $u_t(4) = -\eta$, while $u_t(1) \in \{\pm\eta\}$.

For the case of $i = 2$ (i.e., algorithm (NormSGD)), for every $t < \tau^*(2)$, since $|g(x_t(2))| \geq \varepsilon L_0 |x_1| \geq \gamma$ by (24) and since $x/(x+y)$ is non-decreasing in x on the interval $x \in (0, \infty)$ for any fixed $y \geq 0$,

$$\eta \geq |u_t(2)| = \eta \frac{|g(x_t(2))|}{\gamma + |g(x_t(2))|} \geq \eta \frac{\varepsilon L_0 |x_1|}{\gamma + \varepsilon L_0 |x_1|} = \frac{\eta}{\frac{\gamma}{\varepsilon L_0 |x_1|} + 1} = \lambda_{\text{clip}} \eta \geq \frac{\eta}{2}.$$

Thus, when $t < \tau^*(2)$ and $\xi_{\text{mult},t} = -\varepsilon$, $u_t(2) \geq \lambda_{\text{clip}} \eta = u_t(4)$, and when $t < \tau^*(2)$ and $g_t = (1 + \sigma_1^2/(1+\varepsilon)) L_0 x_t$, $u_t(2) \geq -\eta = \widetilde{u}_t$.

For the case of $i = 3$ (i.e., algorithm (ClippedSGD)), for every $t < \tau^*$, $|g_t| > \gamma$ by (24), which implies that $|u_t| = \eta |g_t| / |g_t| = \eta$. Thus, when $t < \tau^*$ and $g_t = -\varepsilon L_0 x_t$ (notice $g_t > 0$ in this case), $u_t(3) = \eta \geq \lambda_{\text{clip}} \eta = \widetilde{u}_t$, and when $t < \tau^*$ and $g_t = (1 + \sigma_1^2/(1+\varepsilon)) L_0 x_t$, $u_t(3) = -\eta = \widetilde{u}_t$.

Therefore, the claim is established in all three cases, which also concludes the proof. ■

Lemma 38 *Consider the algorithm 4 as defined in Lemma 37. Then, under the assumptions of Lemma 35, we have that, for any $T \geq 1$ and any $t_0 \geq 0$,*

$$\Pr[\tau^*(4) > T] \geq (1-\delta)^{t_0} \left(1 - \sum_{t=t_0+2}^{T} \Pr\left[\frac{1}{t - t_0 - 1} (X_t - \mathbb{E}[X_t]) \leq -(\delta_0 - \delta) - \frac{\delta_0 t_0}{t - t_0 - 1} \right] \right),$$

where $X_t = \sum_{s=t_0+1}^{t-1} \mathbb{1}\{\mathcal{E}_s\}$ is a sum of $t - t_0 - 1$ i.i.d Bernoulli random variables with mean $1 - \delta = 1 - \frac{1}{(1+\sigma_1^2/(1+\varepsilon)^2)}$. $1/\lambda_{\text{clip}} := 1 + \gamma/\varepsilon\sqrt{2\Delta L_0})$ and $\delta_0 = 1/1+1/\lambda_{\text{clip}}$.

Proof Recall the construction of algorithm 4 from Lemma 37. Denote $\mathcal{E}_s = \{\xi_{\text{mult},s} = (1 + \sigma_1^2/(1+\varepsilon))\}$, and recall that $\Pr[\mathcal{E}_s] = \delta = \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2}$. Let us write:

$$\widetilde{N}_{t_1,t_2} = -(x_t(4) - x_1) = \sum_{s=t_1}^{t_2} u_s(4) = \sum_{s=t_1}^{t_2} -\eta \mathbb{1}\{\mathcal{E}_s\} + \lambda_{\text{clip}} \eta \mathbb{1}\{\mathcal{E}_s^c\},$$

as the "net movement" of algorithm 4 to the left of x_{t_1} after $t_2 - t_1 + 1$ time steps. and observe that

$$\mathbb{E}\left[\widetilde{N}_{t_1,t_2}\right] = \sum_{s=t_1}^{t_2} -\eta\delta + \lambda_{\text{clip}}\eta(1-\delta) = \lambda_{\text{clip}}\eta \left(1 - \left(1 + \frac{1}{\lambda_{\text{clip}}}\right)\delta\right)(t_2 - t_1 + 1).$$

Additionally, note that, recalling the definition of $\tau^*(4)$ from Eq. (23),

$$\begin{aligned}\{\tau^*(4) > T\} = \{\forall t \in [2,T] : x_t(4) < x_1\} &= \{\forall t \in [2,T] : -(x_t(4) - x_1) > 0\} \\ &= \left\{\forall t \in [2,T] : \widetilde{N}_{1,t-1} > 0\right\}.\end{aligned}$$

Therefore, we have that, for any $t_0 \geq 0$,

$$\Pr[\tau^*(4) > T] = \Pr\left[\forall t \in [2,T] : \widetilde{N}_{1,t-1} > 0\right] \geq \Pr\left[\left\{\forall t \in [2,T] : \widetilde{N}_{1,t-1} > 0\right\} \cap \bigcap_{s \in [t_0]} \mathcal{E}_s^c\right].$$

Further, since the stochastic gradient of algorithm 4 uses i.i.d multiplicative noise at each round (i.e., the events $\{\mathcal{E}_s\}_{s\in[T]}$ are mutually independent and $\Pr[\mathcal{E}_s] = \delta$ for every s), and since the event $\mathcal{E}_s^c$ implies that $x_{s+1}(i) = x_s(i) - \lambda_{\text{clip}}\eta$ for each algorithm i, we have that for any $t_0 \geq 0$,

$$\begin{aligned}
&\Pr\left[\left\{\forall t \in [2,T] : \widetilde{N}_{1,t-1} > 0\right\} \cap \bigcap_{s\in[t_0]} \mathcal{E}_s^c\right] \\
&= \Pr\left[\left\{\forall t \in [t_0+2,T] : \widetilde{N}_{t_0+1,t-1} > -\lambda_{\text{clip}} t_0 \eta\right\} \cap \bigcap_{s\in[t_0]} \mathcal{E}_s^c\right] \\
&= (1-\delta)^{t_0} \Pr\left[\forall t \in [t_0+2,T] : \widetilde{N}_{t_0+1,t-1} > -\lambda_{\text{clip}} t_0 \eta\right].
\end{aligned}$$

Now, since

$$\begin{aligned}
&\Pr\left[\forall t \in [t_0+2,T] : \widetilde{N}_{t_0+1,t-1} > -\lambda_{\text{clip}} t_0 \eta\right] \\
&= 1 - \Pr\left[\exists t \in [t_0+2,T] : \widetilde{N}_{t_0+1,t-1} \leq -\lambda_{\text{clip}} t_0 \eta\right] \\
&\geq 1 - \sum_{t=t_0+2}^{T} \Pr\left[\widetilde{N}_{t_0+1,t-1} \leq -\lambda_{\text{clip}} t_0 \eta\right],
\end{aligned}$$

it remains only to upper-bound each probability inside of the above summation. To do this, let us denote, for any $t \in [t_0+2, T]$,

$$X_t = \sum_{s=t_0+1}^{t-1} \mathbb{1}\{\mathcal{E}_s^c\} = \frac{1}{(1+\lambda_{\text{clip}})\eta} \widetilde{N}_{t_0+1,t-1} + \frac{t-t_0-1}{1+\lambda_{\text{clip}}}.$$

Thus, X_t is a sum of i.i.d Bernoulli random variables, each with mean $1-\delta = 1 - \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2} > 1/2$ (since, by assumption, $\sigma_1 > (1+\varepsilon)$). We may therefore apply (Hoeffding, 1963, Theorem 1, Eq. (2.2)), denoting $\delta_0 := 1 - \frac{1}{1+\lambda_{\text{clip}}} = \lambda_{\text{clip}}/(1+\lambda_{\text{clip}})$, to obtain:

$$\begin{aligned}
\Pr\left[\widetilde{N}_{t_0+1,t-1} \leq -\lambda_{\text{clip}} t_0 \eta\right] &= \Pr\left[(1+\lambda_{\text{clip}})\eta X_t - \eta(t-t_0-1) \leq -\lambda_{\text{clip}} t_0 \eta\right] \\
&= \Pr\left[\frac{1}{t-t_0-1}(X_t - \mathbb{E}[X_t]) \leq -(\delta_0 - \delta) - \frac{\delta_0 t_0}{t-t_0-1}\right].
\end{aligned}$$

Collecting the above results, we arrive at the claimed lower bound. ■

D.3. Full statement and proof for negative result for (AG-Norm) in the "large σ_1" regime

Lemma 39 (Formal statement of Lemma 34) *Fix any $L_1 > 0$, $x_1 \in \mathbb{R}$, and $\sigma_1 > 1$. Let $T \geq 1$, $\eta > 0$, $\varepsilon \in (0,1)$ and $0 < b_0^2 \leq \varepsilon^2 L_1^2 \exp(2L_1 x_1)$ be arbitrary parameters (possibly dependent on L_1, x_1, and σ_1). Then, there exists a 1-dimensional $(0, (e-1)L_1)$-smooth function such that $F^* = 0$, and a stochastic gradient oracle satisfying Assumptions 3 and 4 with $\sigma_0 = 0$ and the specified σ_1,*

such that, if (AG-Norm) *is run for T time steps using parameters η and b_0^2, then then the resulting iterates $\{x_t\}_{t\in[T]}$ satisfy:*

$$\Pr\left[\min_{t\in[T]}\|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2\right] \geq \left(1 - \frac{1}{1+\frac{\sigma_1^2}{(1+\varepsilon)^2}}\right)^{t_0},$$

where

$$t_0 = \left(1+\sqrt{2}+\frac{\log(T-1)}{2\eta L_1}\right)^2 - 2.$$

In particular, whenever $\eta \geq \alpha/L_1$ for some $\alpha > 0$, and, for any $\delta \in (0,1)$,

$$\sigma_1^2 \geq \frac{1}{\log(1/(1-\delta))}(1+\varepsilon)^2\left(\left(1+\sqrt{2}+\frac{\log(T-1)}{\alpha}\right)^2 - 2\right),$$

then

$$\Pr\left[\min_{t\in[T]}\|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2\right] \geq 1-\delta.$$

Proof Let $F(x) = \exp(L_1 x)$. Notice that, since $\nabla^2 F(x) = L_1 \nabla F(x)$, it follows from Proposition 1 that $F(\cdot)$ is $(0,(e-1)L_1)$-smooth. Clearly $F^* = \inf_{x\in\mathbb{R}}\exp(L_1 x) = 0$. Further, consider the stochastic gradient oracle from Proposition 13, which, for the iterate x_t at time t, first draws an i.i.d sample:

$$\xi_{\text{mult},t} = \begin{cases} -\varepsilon & \text{w.p. } 1-\delta = 1 - \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2} \\ \left(1+\frac{\sigma_1^2}{1+\varepsilon}\right) & \text{w.p. } \delta = \frac{1}{1+\sigma_1^2/(1+\varepsilon)^2}, \end{cases}$$

and $g(x_t) = \xi_{\text{mult},t}\nabla F(x_t)$. As established in Proposition 13, this oracle satisfies Assumptions 3 and 4 with $\sigma_0 = 0$ and the specified $\sigma_1 > 1$.

Let us define, for a parameter $t_0 \geq 1$ to be determined shortly:

$$\mathcal{E}_{\text{nc}} := \{\forall t \in [t_0] : g(x_t) = -\varepsilon\nabla F(x_t)\}.$$

Now, since the noise is sampled i.i.d at each time step, we have that, for any $t_0 \geq 0$:

$$\Pr[\mathcal{E}_{\text{nc}}] = \Pr[\forall t \in [t_0] : \xi_{\text{mult},t} = -\varepsilon] = (1-\delta)^{t_0}.$$

Whenever $\mathcal{E}_{\text{nc}}$ is true, notice that:

$$\begin{aligned}
\nabla F(x_{t_0+1}) = L_1\exp(L_1 x_{t_0+1}) &= L_1\exp\left(L_1 x_1 + L_1\sum_{t=1}^{t_0} x_{t+1} - x_t\right) \\
&= L_1\exp\left(L_1 x_1 + L_1\eta\sum_{t=1}^{t_0}\frac{g(x_t)}{\sqrt{b_0^2+\sum_{s=1}^t\|g(x_s)\|^2}}\right) \\
&= L_1\exp\left(L_1 x_1 + L_1\eta\sum_{t=1}^{t_0}\frac{\varepsilon\nabla F(x_t)}{\sqrt{b_0^2+\sum_{s=1}^t\varepsilon^2\|\nabla F(x_s)\|^2}}\right).
\end{aligned}$$

Now, using the fact that, whenever $\mathcal{E}_{\text{nc}}$ is true, then $\nabla F(x_t) \leq \nabla F(x_{t+1})$ for each $t \in [t_0]$, and assuming $b_0^2 \leq \varepsilon^2 \left\|\nabla F(x_1)\right\|^2$, we can bound

$$\begin{aligned}
\sum_{t=1}^{t_0} \frac{\varepsilon \nabla F(x_t)}{\sqrt{b_0^2 + \sum_{s=1}^{t} \varepsilon^2 \left\|\nabla F(x_s)\right\|^2}} &\geq \sum_{t=1}^{t_0} \frac{\varepsilon^2 \nabla F(x_t)}{\sqrt{\varepsilon \left\|\nabla F(x_1)\right\|^2 + \varepsilon^2 t \left\|\nabla F(x_t)\right\|^2}} \\
&\geq \sum_{t=1}^{t_0} \frac{1}{\sqrt{t+1}} \\
&\geq \int_2^{t_0+2} \frac{1}{\sqrt{t}} dt \\
&= 2(\sqrt{t_0+2} - \sqrt{2}).
\end{aligned}$$

Thus, we conclude that:

$$\nabla F(x_{t_0+1}) \geq L_1 \exp(L_1(x_1 + 2\eta\sqrt{t_0+2} - 2\eta\sqrt{2})).$$

Now, for a parameter $\alpha > 0$ to be determined shortly, let us define:

$$\tau_0 = \min\left\{t \geq t_0 : \nabla F(x_{t+1}) \leq \nabla F(x_{t_0+1}) \exp(-L_1\eta\alpha)\right\},$$

and let, for each $i \geq 0$,

$$\tau_{i+1} = \min\left\{t \geq \tau_i : \nabla F(x_{t+1}) < \nabla F(x_{\tau_i})\right\}.$$

Notice that, by construction, $\xi_{\text{mult},\tau_i} = 1 + \sigma_1^2/(1+\varepsilon)$ for every $i \geq 0$. Further, $x_{\tau_i+1} \leq x_{\tau_{i+1}}$ since τ_{i+1} is the first time after τ_i satisfying $\nabla F(x_{\tau_{i+1}+1}) < \nabla F(x_{\tau_i+1})$, or equivalently, $x_{\tau_{i+1}+1} < x_{\tau_i+1}$. This implies that

$$\begin{aligned}
x_{\tau_i+1} &= x_{\tau_0+1} + \sum_{j=0}^{i-1} x_{\tau_{j+1}+1} - x_{\tau_j+1} \geq x_{\tau_0+1} + \sum_{j=0}^{i-1} x_{\tau_{j+1}+1} - x_{\tau_{j+1}} \\
&= x_{\tau_0+1} - \sum_{j=0}^{i-1} \frac{\eta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\nabla F(x_{\tau_{j+1}})}{\sqrt{b_0^2 + \sum_{s=1}^{t} g_t^2}} \geq x_{\tau_0+1} - \sum_{j=0}^{i-1} \frac{\eta\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)\nabla F(x_{\tau_{j+1}})}{\sqrt{\left(1 + \frac{\sigma_1^2}{1+\varepsilon}\right)^2 \left\|\nabla F(x_{t_0+1})\right\|^2}}.
\end{aligned}$$

Now, notice that:

$$\begin{aligned}
\nabla F(x_{\tau_{j+1}}) &= L_1 \exp(L_1 x_{\tau_{j+1}+1} + L_1(x_{\tau_{j+1}} - x_{\tau_{j+1}+1})) \\
&= \nabla F(x_{\tau_{j+1}+1}) \exp(L_1(x_{\tau_{j+1}} - x_{\tau_{j+1}+1})) \\
&< \nabla F(x_{\tau_0+1}) \exp(L_1(x_{\tau_{j+1}} - x_{\tau_{j+1}+1})) \\
&\leq \nabla F(x_{t_0+1}) \exp(L_1(x_{\tau_{j+1}} - x_{\tau_{j+1}+1} - \eta\alpha)) \\
&\leq \nabla F(x_{t_0+1}) \exp(-\eta L_1(\alpha - 1)),
\end{aligned}$$

from which we obtain the bound:

$$x_{\tau_i+1} \geq x_{\tau_0+1} - \eta \sum_{j=0}^{i-1} \exp(-\eta L_1(\alpha-1)) = x_{\tau_0+1} - i\eta \exp(-\eta L_1(\alpha-1)).$$

Now, by construction of the τ_i, we have that, assuming $\mathcal{E}_{\text{nc}}$ is true, then

$$\begin{aligned}\min_{t\in[T]} \|\nabla F(x_t)\|^2 &= \min_{t\in[t_0+2,T]} \min\left\{\|\nabla F(x_1)\|^2, \|\nabla F(x_t)\|^2\right\} \\ &\geq \min_{i\in[0,T-1]} \min\left\{\|\nabla F(x_1)\|^2, \|\nabla F(x_{\tau_i+1})\|^2\right\}.\end{aligned}$$

Thus, to ensure that $\min_{t\in[T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2$, it suffices to have that, for every $i < T$, $\|\nabla F(x_{\tau_i})\|^2 \geq \|\nabla F(x_1)\|^2$. Now, notice that

$$\begin{aligned}\nabla F(x_{\tau_i+1}) &= L_1 \exp(L_1(x_{\tau_i+1})) \\ &\geq L_1 \exp(L_1(x_{\tau_0+1}) - i\eta \exp(-\eta L_1(\alpha-1))) \\ &= \nabla F(x_{\tau_0+1}) \exp(-iL_1\eta \exp(-\eta L_1(\alpha-1))) \\ &= \nabla F(x_{\tau_0}) \exp(L_1(x_{\tau_0+1} - x_{\tau_0}) - iL_1\eta \exp(-\eta L_1(\alpha-1))) \\ &> \nabla F(x_{t_0+1}) \exp(-L_1\eta\alpha + L_1(x_{\tau_0+1} - x_{\tau_0}) - iL_1\eta \exp(-\eta L_1(\alpha-1))) \\ &\geq \nabla F(x_1) \exp(2\eta L_1(\sqrt{t_0+2} - \sqrt{2}) - L_1\eta(\alpha+1) - iL_1\eta \exp(-\eta L_1(\alpha-1))).\end{aligned}$$

Thus, it suffices to establish conditions under which

$$2\sqrt{t_0+2} \geq 2\sqrt{2} + \alpha + 1 + i\exp(-L_1\eta(\alpha-1)).$$

Thus, if we choose $\alpha = {\log(T-1)}/{\eta L_1}$, then it suffices to take:

$$t_0 = \left(1 + \sqrt{2} + \frac{\log(T-1)}{2\eta L_1}\right)^2 - 2,$$

in which case $\min_{t\in[T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2$ under $\mathcal{E}_{\text{nc}}$. Hence,

$$\begin{aligned}\Pr\left[\min_{t\in[T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2\right] &\geq \Pr\left[\mathcal{E}_{\text{nc}}\right] \\ &= \left(1 - \frac{1}{1 + \frac{\sigma_1^2}{1+\varepsilon}}\right)^{t_0}.\end{aligned}$$

In particular, using the fact that $1 - x > \exp({-x}/{(1-x)})$ for $x < 1$, it follows that:

$$\begin{aligned}\Pr\left[\min_{t\in[T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2\right] &\geq \exp\left(-\frac{(1+\varepsilon)^2 t_0}{\sigma_1^2}\right) \\ &\geq \exp\left(-\frac{(1+\varepsilon)^2\left(\left(1+\sqrt{2}+\frac{\log(T-1)}{\eta L_1}\right)^2 - 2\right)}{\sigma_1^2}\right)\end{aligned}$$

Hence, as long as, for some $\delta \in (0,1)$,

$$\sigma_1^2 \geq \frac{1}{\log({1}/{(1-\delta)})}(1+\varepsilon)^2\left(\left(1+\sqrt{2}+\frac{\log(T-1)}{\eta L_1}\right)^2 - 2\right),$$

then $\Pr\left[\min_{t\in[T]} \|\nabla F(x_t)\|^2 = \|\nabla F(x_1)\|^2\right] \geq 1 - \delta$. ■