# SGD with AdaGrad Stepsizes: Full Adaptivity with High Probability to Unknown Parameters, Unbounded Gradients and Affine Variance

**Amit Attia** [1] **Tomer Koren** [1 2]

## Abstract

We study Stochastic Gradient Descent with AdaGrad stepsizes: a popular adaptive (self-tuning) method for first-order stochastic optimization. Despite being well studied, existing analyses of this method suffer from various shortcomings: they either assume some knowledge of the problem parameters, impose strong global Lipschitz conditions, or fail to give bounds that hold with high probability. We provide a comprehensive analysis of this basic method without any of these limitations, in both the convex and non-convex (smooth) cases, that additionally supports a general "affine variance" noise model and provides sharp rates of convergence in both the low-noise and high-noise regimes.

## 1. Introduction

Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951) is the de facto standard optimization method in modern machine learning. SGD updates take the following form;

$$w_{t+1} = w_t - \eta_t g_t,$$

starting from an arbitrary point $w_1$, where $\eta_t$ is the stepsize at time $t$ and $g_t$ is a stochastic gradient at $w_t$. The sequence of stepsizes $\eta_1, \eta_2, \ldots$ is notoriously hard to tune in machine learning applications (e.g., Bottou, 2012; Schaul et al., 2013), especially when the problem parameters (e.g., smoothness or Lipschitz constant of the objective) or estimates thereof are unknown. This has popularized the study of *adaptive gradient methods* (e.g., Duchi et al., 2011; Kingma & Ba, 2015; Reddi et al., 2018; Tran et al., 2019; Alacaoglu et al., 2020) where stepsizes are computed adaptively based on magnitudes of previous stochastic gradients.

[1]Blavatnik School of Computer Science, Tel Aviv University [2]Google Research Tel Aviv. Correspondence to: Amit Attia <amitattia@mail.tau.ac.il>.

*Proceedings of the 40$^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Although their non-adaptive counterparts are well studied, optimization theory of adaptive methods is still lacking and attract much interest, particularly in the smooth and non-convex setting. The different types of analyses performed by previous works in this setting revolve around the assumptions taken in order to establish convergence which includes the type of noise (bounded variance, sub-Gaussian, bounded support, affine variance, etc.), whether the true gradients or the stochastic gradients are uniformly bounded (Ward et al., 2019; Kavis et al., 2022), and also whether to use a step-size that does not depends on the current stochastic gradient (Li & Orabona, 2019; 2020) or to handle the bias introduced by such a dependency (Ward et al., 2019; Faw et al., 2022).

We focus on perhaps the simplest and most well-studied adaptive optimization method: SGD with AdaGrad-type stepsizes, which we refer to as AdaSGD (also known in the literature as AdaGrad-Norm; e.g., Ward et al. 2019; Faw et al. 2022). For parameters $\eta, \gamma > 0$, AdaSGD performs SGD updates with stepsizes

$$\eta_t = \frac{\eta}{\sqrt{\gamma^2 + \sum_{s=1}^{t} \|g_s\|^2}},$$

where $g_t$ is a stochastic gradient at $w_t$.

Although numerous convergence results for smooth AdaSGD exist in the literature (mostly in the non-convex case, except for Li & Orabona (2019) discussed below), they all fall short of at least one important aspect. First, many of the existing analyses require a strong global Lipschitz assumption on the objective (Ward et al., 2019; Zhou et al., 2018; Kavis et al., 2022) which does not hold even for simple quadratic functions in one dimension, and one that is not required for the analysis of analogous non-adaptive methods. Second, several results assume partial knowledge of the problem parameters, such as the smoothness parameter (e.g., Li & Orabona, 2019; 2020), and thus in a sense do not fully establish the desired adaptivity of AdaSGD. A summary of the various results for SGD with AdaGrad stepsizes is presented in Table 1.

To our knowledge, only Faw et al. (2022) do not impose neither of the above limiting assumptions, but their bounds do not hold with high probability: namely, they give convergence bounds with a polynomial dependence on the prob-

ability margin (granted, under a somewhat weaker tail assumption on the gradient noise). On the other hand, their results apply in a significantly more challenging setting with an "affine variance" noise model, where the variance of the gradient noise is not uniformly bounded and may scale with the squared norm of the true gradient.

### 1.1. Our contributions

We provide a comprehensive analysis of AdaSGD that does not suffer from any of the former issues: for a $\beta$-smooth objective and under a general affine variance noise model, we establish high probability convergence guarantees in both the non-convex and convex regimes, without any knowledge of the problem parameters and without imposing a global Lipschitz condition on the objective nor a bound on the distance to a minimizer. Further, our analysis is considerably simpler than previous analyses of AdaSGD in similar generality (e.g., Faw et al., 2022).

In more detail, we consider a stochastic first-order model with affine variance noise, similar to the one of Faw et al. (2022); given a point $w \in \mathbb{R}^d$ we assume that the gradient oracle returns a random vector $g(w)$ such that $\mathbb{E}[g(w) \mid w] = \nabla f(w)$ and, with probability one, satisfies[1]

$$\|g(w) - \nabla f(w)\|^2 \le \sigma_0^2 + \sigma_1^2 \|\nabla f(w)\|^2,$$

where $\sigma_0, \sigma_1 \ge 0$ are noise parameters, unknown to the algorithm. Our main results in this setup are as follows:

- Our first main result establishes a high probability convergence guarantee in the general $\beta$-smooth (not necessarily convex) setting. Given a $\beta$-smooth objective $f$, we show that for $T$-steps AdaSGD, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|^2 \le C_1 \frac{\mathrm{poly}(\log \frac{T}{\delta})}{T} + C_2 \frac{\sigma_0 \,\mathrm{poly}(\log \frac{T}{\delta})}{\sqrt{T}},$$

  where $C_1$ and $C_2$ are quantities that depend polynomially on the problem parameters ($\beta, \sigma_0, \sigma_1, \eta, \gamma$; the dependence on $1/\gamma$ is in fact logarithmic). The formal statement is given in Theorem 1.

- Our second main result is a convergence guarantee for AdaSGD in the same unconstrained optimization setup but with an additional convexity assumption. Given a $\beta$-smooth and *convex* objective $f$, we show that for $T$-steps AdaSGD, with probability at least $1-\delta$, it holds that

$$f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - f^\star \le C_3 \frac{\mathrm{poly}(\log \frac{T}{\delta})}{T} + C_4 \frac{\sigma_0 \,\mathrm{poly}(\log \frac{T}{\delta})}{\sqrt{T}},$$

  where $f^\star = \min_{w \in \mathbb{R}^d} f(w)$ and $C_3$ and $C_4$ are again quantities that depend polynomially on the problem parameters ($\beta, \sigma_0, \sigma_1, \eta, \gamma$; here the dependence on $1/\gamma$ is polynomial). The formal statement is provided in Theorem 2.

[1]While our analysis assumes for simplicity affine noise bounded with probability one, we provide a general reduction to allow for sub-Gaussian noise with bounded affine variance; see Appendix C for details.

These convergence bounds match the rates of the form $O(1/T + \sigma_0/\sqrt{T})$ for *non-adaptive, properly tuned SGD* in the convex (Lan, 2012) and non-convex cases (Ghadimi & Lan, 2013) cases, up to problem dependent constants and logarithmic factors. However, we emphasize again that our results hold true, with high probability, *without* assuming uniformly bounded gradients and *without* requiring any knowledge of the problem parameters ($\beta, \sigma_0, \sigma_1$). To our knowledge, in the non-convex case our result is the first convergence guarantee to hold with high probability (with a proper poly-log dependence on the probability margin) without requiring a global Lipschitz assumption—and in fact, it holds even in a more general affine variance noise model. Similarly, as far as we are aware, our result in the convex case constitutes the first (high probability) analysis of adaptive SGD updates that applies in an unconstrained optimization setup, i.e., without restricting the updates to a domain of bounded (and a-priori known) diameter using projections.

Further, our bounds maintain a remarkable property of the non-adaptive bounds for smooth SGD (Lan, 2012; Ghadimi & Lan, 2013): as can be seen from the bounds above, they exhibit optimal rates (in the context of non-accelerated gradient methods) in both the high-noise and low-noise regimes, in the sense that they improve from the traditional stochastic $1/\sqrt{T}$ rate to a fast $1/T$ rate, up to logarithmic factors and problem-dependent constants, when the additive noise parameter $\sigma_0$ is sufficiently small: namely, when $\sigma_0 = O(1/\sqrt{T})$. In contrast to the non-adaptive bounds, however, this interpolation between different rates occurs automatically, without knowledge of the underlying parameters and their relevant regime. Unlike prior work on the affine noise model, our rate-interpolation takes place even when the affine term is large, i.e., when $\sigma_1 \gg 0$; for comparison, Faw et al. (2022) required that $\sigma_1 = O(1/\sqrt{T})$ for drawing a similar conclusion. For completeness, we also provide a high probability analysis of non-adaptive, smooth SGD (with known parameters) under the affine noise model, complementing the result in expectation of Bottou et al. (2018, Theorem 4.8); see details in Appendix E.

Finally, it is worth highlighting another remarkable adaptivity property of AdaSGD, demonstrated by our results: the algorithm is able to automatically adapt to the optimization problem at hand and the same step-size schedule simultaneously achieves the optimal rates of convergence of (non-accelerated) SGD in both the convex and non-convex smooth (even unconstrained) optimization settings.

Table 1: Convergence results for non-convex SGD with AdaGrad stepsizes under different assumptions. For simplicity, the rate column consider the dependence in $T$ (number of steps) and $\sigma$ (noise variance), ignoring other parameters and logarithmic factors. The "high probability" column indicates results with poly-logarithmic (rather than polynomial) dependence in the probability margin. The "convex guarantee" column indicates an additional analysis for the convex case.

| | ADAPTIVE STEPSIZE | UNKNOWN SMOOTHNESS | HIGH PROBABILITY | UNBOUNDED GRADIENTS | AFFINE NOISE | CONVEX GUARANTEE | NON-CONVEX RATE |
|---|---|---|---|---|---|---|---|
| Ghadimi & Lan (2013)[a] | | | ✓ | ✓ | | | $\frac{1+\sigma^2}{T} + \frac{\sigma}{\sqrt{T}}$ |
| Bottou et al. (2018) | | | | ✓ | ✓ | | $\frac{1}{T} + \frac{\sigma}{\sqrt{T}}$ |
| **This paper** (Thm. 5) | | | ✓ | ✓ | ✓ | ✓[b] | $\frac{1+\sigma^2}{T} + \frac{\sigma}{\sqrt{T}}$ |
| Li & Orabona (2019) | ✓ | | | ✓ | | ✓ | $\frac{1}{T} + \frac{\sigma}{\sqrt{T}}$ |
| Ward et al. (2019) | ✓ | ✓ | | | | | $\frac{1}{T} + \frac{\sigma(1+\sigma)}{\sqrt{T}}$ |
| Kavis et al. (2022) | ✓ | ✓ | ✓ | | | | $\frac{1+\sigma^4}{T} + \frac{\sigma(1+\sigma^2)}{\sqrt{T}}$ |
| Faw et al. (2022) | ✓ | ✓ | | ✓ | ✓ | | $\frac{1}{T} + \frac{\sigma(1+\sigma)}{\sqrt{T}}$ |
| Liu et al. (2023)[c] | ✓ | ✓ | ✓ | ✓ | | | $\frac{1+\sigma^5}{T} + \frac{\sigma(1+\sigma^3)}{\sqrt{T}}$ |
| **This paper** (Thms. 1 & 2) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | $\frac{1}{T} + \frac{\sigma(1+\sigma)}{\sqrt{T}}$ |

[a] The high probability result is obtained by amplifying the success probability using multiple runs of SGD.
[b] Though not formalized in this paper, a result for fixed-stepsize SGD in the convex case can be extracted from our analysis in the adaptive case.
[c] This manuscript has appeared on arXiv after the publication of the present work.

### 1.2. Analysis overview

Let us highlight some of the key challenges in analyzing SGD with adaptive stepsizes and outline our approach for overcoming them. We begin with the non-convex smooth case, which has been the focus of prior work.

**Non-convex regret analysis.** The starting point of our analysis is a "non-convex regret bound" for AdaSGD, similar to Kavis et al. (2022); for a $\beta$-smooth objective $f$ it can be shown (see Lemma 1 for details and a proof) that

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t \leq \frac{\gamma \Delta_1}{\eta} + \left(\frac{2\bar{\Delta}_T}{\eta} + \eta\beta\right)\sqrt{\sum_{t=1}^{T} \|g_t\|^2},$$

where $\Delta_1 = f(w_1) - f^\star$ and $\bar{\Delta}_T = \max_{t \leq T} f(w_t) - f^\star$.

The challenging term to handle in this analysis turns out to be the one that involves $\bar{\Delta}_T$. Indeed, assume for example a deterministic bound $\bar{\Delta}_T \leq F$ (e.g., this is the case whenever $f$ is upper bounded uniformly); taking expectations of both sides, we obtain

$$\sum_{t=1}^{T} \mathbb{E}\|\nabla f(w_t)\|^2 \leq \frac{\gamma \Delta_1}{\eta} + \left(\frac{2F}{\eta} + \eta\beta\right)\sqrt{\sum_{t=1}^{T} \mathbb{E}\|g_t\|^2}.$$

Then, using $\|g_t\|^2 \leq 2\|\nabla f(w_t) - g_t\|^2 + 2\|\nabla f(w_t)\|^2$ and our assumption on the noise yields the bound

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|^2\right] \leq \frac{2\gamma\Delta_1}{\eta T} + (1+\sigma_1^2)\left(\frac{2F}{\eta} + \eta\beta\right)^2 \frac{1}{T} + \left(\frac{2F}{\eta} + \eta\beta\right)\sigma_0\sqrt{\frac{8}{T}},$$

which concludes an in-expectation analysis.

**Bounding $\bar{\Delta}_T$ w.h.p.** The issue, however, is that $\bar{\Delta}_T$ is a random variable which is not bounded a-priori. Due to the probabilistic dependence between the different factors in the regret bound, our approach proceeds by establishing a bound over $\bar{\Delta}_T$ that holds with high probability (rather than in expectation). Due to smoothness, we can bound $\bar{\Delta}_T$ via

$$f(w_{t+1}) \leq f(w_1) - \sum_{s=1}^{t} \eta_s \nabla f(w_s) \cdot g_s + \frac{\beta}{2}\sum_{s=1}^{t} \eta_s^2 \|g_s\|^2.$$

The second summation in the bound is straightforward to handle and can be bounded (deterministically) by a logarithmic term (see Lemma 6 for details). The first summation is more problematic, again due to (conditional) probabilistic dependencies involving the stepsize $\eta_s$, which is crucially a random variable that depends on stochastic gradient $g_s$.

**Decorrelated stepsizes.** To break the correlation between $\eta_s$ and $g_s$ in the second term above, Kavis et al. (2022) replace $\eta_s$ with $\eta_{s-1}$. Then, for bounding the difference between $\eta_s$ and $\eta_{s-1}$ they assumed a *global Lipschitz assumption* which limits the effect of each individual gradient on the stepsize $\eta_s$. To avoid this limiting assumption, we use a different "decorrelated" proxy for the stepsize, defined by replacing $\|g_t\|^2$ with $(1+\sigma_1^2)\|\nabla f(w_t)\|^2 + \sigma_0^2$; this was introduced first by Ward et al. (2019) and then modified by Faw et al. (2022) to allow for affine-variance noise. We can then decompose using the proxy,

$$-\eta_s \nabla f(w_s) \cdot g_s = -\tilde{\eta}_s \|\nabla f(w_s)\|^2 + \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) + (\tilde{\eta}_s - \eta_s)\nabla f(w_s) \cdot g_s,$$

and bound the cumulative sum of each of the terms.

**Concentration with affine variance.** The challenging term to bound is the martingale

$$\sum_{s=1}^{t} \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s).$$

Bounding the martingale is challenging because the magnitude $\|\nabla f(w_s) - g_s\|$ depends on the previous steps with affine-variance noise. The key observation is that since we wish to establish a bound for $f(w_{t+1}) - f^\star$ for all $t$, and thanks to the well-known property of smooth functions that $\|\nabla f(w)\|^2 \leq 2\beta(f(w) - f^\star)$, it is suffice to perform an inductive argument and use the bound for the previous steps to bound the affine noise terms. As the noise bound in a given step is in itself a random variable that depends on previous stochastic gradients, we appeal to a Bernstein-type martingale concentration inequality (specifically, we use one due to Li & Orabona 2020) and a careful analysis yields a bound on $-\sum_{s=1}^{t} \eta_s \nabla f(w_s) \cdot g_s$ and establishes a bound on $\bar{\Delta}_T$ without the need for a global Lipschitz assumption.

**Additional challenges in the convex case.** We follow a similar approach using a regret bound and a bound of $\|w_t - w^\star\|^2$ which holds for all $t \leq T$ with high probability. In the convex case we need to perform a similar decomposition to bound $-\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w^\star)$ where $w^\star$ is a minimizer of $f$. Using the proxy stepsize sequence $\tilde{\eta}_t$ does not work. Instead, we employ a slightly different decorrelated stepsize sequence $(\hat{\eta}_s)_s$ that (A) suffices for bounding $\sum_{s=1}^{t} (\hat{\eta}_s - \eta_s) g_s \cdot (w_s - w^\star)$ and (B) keeps the norm of the zero-mean vector $\hat{\eta}_s(\nabla f(w_s) - g_s)$ bounded with probability one, which is crucial for the martingale analysis. The martingale lemma of Li & Orabona (2020) is too loose for bounding $\sum_{s=1}^{t} \hat{\eta}_s(\nabla f(w_s) - g_s) \cdot (w_s - w^\star)$ and we rely instead on the martingale analysis technique of Carmon & Hinder (2022) which uses an empirical Bernstein inequality introduced by Howard et al. (2021). The empirical Bernstein is necessary in order to tightly bound the martingale and establish the bound on $\|w_t - w^\star\|^2$.

### 1.3. Additional related work

**SGD for non-convex smooth optimization.** The first to analyze SGD in the smooth non-convex setting and obtain tight convergence bounds were Ghadimi & Lan (2013). They established that a *properly tuned* $T$-steps SGD obtain a rate of $O(1/T + \sigma/\sqrt{T})$ assuming uniformly bounded variance of $\sigma^2$. A tight lower bound is given by Arjevani et al. (2022). Bottou et al. (2018, Theorem 4.8) extended the convergence result for the affine variance noise model.

**Adaptive methods for non-convex smooth optimization.** An extensive line of work tackles convergence guarantees in expectation.[2] Ward et al. (2019) gave a rate of $\widetilde{O}(1/\sqrt{T})$ for $T$-steps Adaptive SGD, assuming uniformly bounded gradients. In parallel, Li & Orabona (2019) obtained a rate of $\widetilde{O}(1/\sqrt{T})$ without assuming uniformly bounded gradients, assuming instead knowledge of the smoothness parameter and using a step-size that does not depend on the current stochastic gradient. Li & Orabona (2019) also obtain a convergence guarantee for the smooth and convex case, again assuming knowledge of the smoothness. A recent work by Faw et al. (2022) assumed affine variance noise and achieved the desired $\widetilde{O}(1/\sqrt{T})$ in the noisy regime without assuming uniformly bounded gradients.

**High probability convergence bounds.** Given knowledge of the smoothness and sub-Gaussian noise bounds, Ghadimi & Lan (2013) established convergence with high probability for non-adaptive SGD with properly tuned step-size. Few high probability results of adaptive methods exist. Zhou et al. (2018) and Li & Orabona (2020) established a high probability convergence result for AdaGrad and delayed AdaGrad respectively, assuming sub-Gaussian noise and knowledge of the smoothness to tune the step-size. Kavis et al. (2022) obtained a high probability convergence for Adaptive SGD, assuming uniformly bounded gradients and either sub-Gaussian noise or uniformly bounded stochastic gradients. Assuming heavy-tailed noise, Cutkosky & Mehta (2021) obtained convergence with high probability for normalized SGD with clipping and momentum with a tuned step-size according to a moment bound of the noise.

**Unconstrained stochastic convex optimization.** Also related to our work are Harvey et al. (2019) and Carmon & Hinder (2022) who focused on the (non-smooth) unconstrained and convex setting. Harvey et al. (2019) established convergence with high probability of SGD using a generalized freedman's inequality. Carmon & Hinder (2022) devised a stochastic parameter-free method which obtain high probability convergence guarantee and their method also obtain optimal rate for *noiseless* convex and smooth objective. Both papers used a careful martingale analysis in order to bound the distance of the iterates from a minimizer, a desired property for unconstrained stochastic optimization which we also establish for AdaSGD.

**Parameter-free methods for online optimization.** The closely related literature on online optimization includes a plethora of adaptive methods. Recent work focuses on adaptivity to the norm of the comparator (e.g., Orabona & Pál, 2016; Cutkosky & Orabona, 2018; Mhammedi & Koolen, 2020) and to the scale of the objective (e.g., Orabona & Pál, 2018; Mhammedi & Koolen, 2020).

[2] We include results with polynomial dependence in the probability margin as they can be derived using Markov's inequality.

## 2. Preliminaries

We consider unconstrained smooth optimization in $d$-dimensional Euclidean space $\mathbb{R}^d$ with the $\ell_2$ norm, denoted $\|\cdot\|$. A function $f$ is said to be $\beta$-smooth if for all $x, y \in \mathbb{R}^d$ it holds that $\|\nabla f(x) - \nabla f(y)\| \le \beta\|x - y\|$. This implies, in particular, that for all $x, y \in \mathbb{R}^d$, the following bound is true: $f(y) \le f(x) + \nabla f(x) \cdot (y - x) + \frac{1}{2}\beta\|y - x\|^2$. We additionally assume that $f$ is lower bounded over $\mathbb{R}^d$, and admits a minimum $f^\star = \min_{w \in \mathbb{R}^d} f(w)$. We assume a standard stochastic first-order model, in which we are provided with an oracle $g(w)$ where, given a point $w \in \mathbb{R}^d$ returns a random vector $g(w) \in \mathbb{R}^d$ such that $\mathbb{E}[g(w) \mid w] = \nabla f(w)$.

**Affine noise model.** We assume a general noise model, referred to in the literature as the "affine variance noise" (e.g., Faw et al. (2022) and references therein). In this model, the gradient oracle $g$ satisfies

$$\forall\, w \in \mathbb{R}^d : \quad \|g(w) - \nabla f(w)\|^2 \le \sigma_0^2 + \sigma_1^2\|\nabla f(w)\|^2$$

(deterministically), where $\sigma_0, \sigma_1 \ge 0$ are noise parameters. For simplicity, we will assume throughout a deterministic bound on the noise magnitude, as stated above; in Appendix C, through a generic reduction, we extend our arguments to apply in a more general sub-Gaussian noise model with affine variance.

The affine noise model permits the noise magnitude to grow as the norm of the true gradient increases (without the affine term, i.e., assuming $\sigma_1 = 0$, the gradients are effectively "noiseless" for gradients of large magnitude). We refer to Faw et al. (2022) for an additional discussion of this model.

**Adaptive stepsize Stochastic Gradient Descent.** Starting from an arbitrary point $w_1 \in \mathbb{R}^d$, for parameters $\eta, \gamma > 0$, the update rule of AdaSGD is

$$w_{t+1} = w_t - \eta_t g_t,$$

where $g_t = g(w_t)$ is a stochastic gradient at $w_t$ and the step-size $\eta_t$ is given by

$$\forall\, 1 \le t \le T : \qquad \eta_t = \frac{\eta}{\sqrt{\gamma^2 + \sum_{s=1}^t \|g_s\|^2}}. \tag{1}$$

**Additional notation.** Throughout, we use $\mathbb{E}_{t-1}[\cdot] \triangleq \mathbb{E}[\cdot \mid g_1, \ldots, g_{t-1}]$ to denote the conditional expectation with respect to the stochastic gradients queried up to (not including) step $t$. We use log to denote the base 2 logarithm.

## 3. Analysis in the general non-convex case

First, we focus on the general smooth case without convexity assumptions. Our main result of this section is the following convergence bound for AdaSGD.

**Theorem 1.** *Assume that $f$ is $\beta$-smooth. Then for AdaSGD, for any $\delta \in (0, 1/3)$, it holds with probability at least $1 - \delta$ that*

$$\text{(i)}\ \ \frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|^2 \le O\Big(\tfrac{2F}{\eta} + \eta\beta\Big)\frac{\sigma_0}{\sqrt{T}} + O\Big(\tfrac{\gamma}{\eta}\Delta_1 + (1 + \sigma_1^2 \log\tfrac{1}{\delta})\big(\tfrac{2F}{\eta} + \eta\beta\big)^2 + \sigma_0^2 \log\tfrac{1}{\delta}\Big)\frac{1}{T};$$

*(ii) and $f(w_t) - f^\star \le F$ for all $1 \le t \le T + 1$,*

*where $\Delta_1 = f(w_1) - f^\star$ and*

$$F = 2\Delta_1 + \big(3\log\tfrac{T}{\delta} + 4C_1\big)\eta\sigma_0 + \big(9\log^2\tfrac{T}{\delta} + 16C_1^2\big)\eta^2\beta\sigma_1^2 + \eta^2\beta C_1, \tag{2}$$

$$C_1 = \log\left(1 + \frac{2\sigma_0^2 T + 8(1 + \sigma_1^2)(\eta^2\beta^2 T^3 + \beta\Delta_1 T)}{\gamma^2}\right). \tag{3}$$

As stated, the bound in the theorem holds for bounded affine noise; for the precise bound in the more general sub-Gaussian case, see Theorem 3 in Appendix C.

In addition to the features of this convergence result already highlighted in the introduction, we also note the high probability guarantee it provides that the maximal optimality gap $f(w_t) - f^\star$ remains $\widetilde{O}(1)$ along the trajectory of AdaSGD. Comparing the result to the high probability analysis when the parameters are known (Appendix E), the main difference is the bound of $f(w_t) - f^\star$ which is $O(\Delta_1)$ when the parameters are known (for large enough $T$), while the bound of Theorem 1 has additional terms that depend on $\sigma_0$ and $\sigma_1$. We remark that the prior results of Faw et al. (2022) in this setting include a somewhat worse high-order noise term, of order $\sigma_1(\sigma_0^2 + \sigma_1^{12})/\sqrt{T}$ in the noisy regime (for $\sigma_1 = \omega(1)$).

### 3.1. Proof of Theorem 1

Henceforth we use the following notations: let $\Delta_t \triangleq f(w_t) - f^\star$ and $\bar{\Delta}_t \triangleq \max_{s \le t} f(w_s) - f^\star$, for all $t$. We also denote $G_t \triangleq (\gamma^2 + \sum_{s=1}^t \|g_s\|^2)^{1/2}$, thus, $\eta_t = \eta/G_t$. As $\eta_t$ and $g_t$ are correlated random variables (indeed, $\eta_t$ depends on the norm of $g_t$), we define the following "decorrelated stepsizes,"

$$\forall\, 1 \le t \le T : \tilde{\eta}_t \triangleq \frac{\eta}{\sqrt{G_{t-1}^2 + (1 + \sigma_1^2)\|\nabla f(w_t)\|^2 + \sigma_0^2}}.$$

We begin by stating several lemmas that we use to prove Theorem 1 (proofs are given later in the section). The first lemma states a "non-convex regret bound."

**Lemma 1.** *If $f$ is $\beta$-smooth then for AdaSGD we have*

$$\sum_{t=1}^T \nabla f(w_t) \cdot g_t \le \frac{\gamma\Delta_1}{\eta} + \left(\frac{2\bar{\Delta}_T}{\eta} + \eta\beta\right)\sqrt{\sum_{t=1}^T \|g_t\|^2}.$$

Note that the $\bar{\Delta}_T$ term in the right hand-side of the bound is a random variable. In the noiseless setting with known smoothness and properly tuned stepsize, we know that $\bar{\Delta}_T = \Delta_1$ since $\Delta_{t+1} \le \Delta_t$ for all $t$ and proving convergence would be much simplified. The next lemma bounds $\bar{\Delta}_T = \max_{s \le t} f(w_s) - f^\star$ with high probability in the noisy case.

**Lemma 2.** *With probability at least $1-\delta$, it holds that $\bar{\Delta}_{T+1} \le F$ (with $F$ defined in Equation (2)).*

In order to translate the regret analysis of Lemma 1 to a high probability bound we need a bound on the sum of differences between $\|\nabla f(w_t)\|^2$ and $g_t \cdot \nabla f(w_t)$ which we obtain in the following lemma (see Appendix A for a proof).

**Lemma 3.** *with probability at least $1-2\delta$, it holds that*

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot (\nabla f(w_t) - g_t) \le \frac{1}{4}\sum_{t=1}^{T} \|\nabla f(w_t)\|^2 + 3(\sigma_0^2 + 2\beta\sigma_1^2 F)\log\tfrac{1}{\delta},$$

*with $F$ given in Equation (2).*

Using the lemmas above we can now prove our main result.

*Proof of Theorem 1.* All arguments below follow given the high probability events of Lemmas 2 and 3. Hence, using a union bound, they follow with probability at least $1-3\delta$. We denote $C_2 = 2F/\eta + \eta\beta$ to simplify notation. From Lemmas 1 and 2,

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t \le \frac{\gamma\Delta_1}{\eta} + C_2\sqrt{\sum_{t=1}^{T}\|g_t\|^2}.$$

Observe that

$$\begin{aligned}\|g_t\|^2 &\le 2\|\nabla f(w_t)\|^2 + 2\|g_t - \nabla f(w_t)\|^2 \\ &\le 2(1+\sigma_1^2)\|\nabla f(w_t)\|^2 + 2\sigma_0^2.\end{aligned}$$

Hence, using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$,

$$\sqrt{\sum_{t=1}^{T}\|g_t\|^2} \le \sqrt{2\sum_{t=1}^{T}(1+\sigma_1^2)\|\nabla f(w_t)\|^2} + \sigma_0\sqrt{2T}.$$

By the elementary fact $ab \le a^2/2 + b^2/2$ with $a = \sqrt{\frac{1}{2}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2}$ and $b = 2C_2\sqrt{1+\sigma_1^2}$,

$$C_2\sqrt{2(1+\sigma_1^2)\sum_{t=1}^{T}\|\nabla f(w_t)\|^2} \le \frac{1}{4}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 + 2(1+\sigma_1^2)C_2^2.$$

Hence,

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t \le \frac{\gamma\Delta_1}{\eta} + 2(1+\sigma_1^2)C_2^2 + C_2\sigma_0\sqrt{2T} + \frac{1}{4}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2.$$

Summing the above with Lemma 3 and rearranging,

$$\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 \le \frac{2\gamma\Delta_1}{\eta} + 4(1+\sigma_1^2)C_2^2 + 6(\sigma_0^2 + 2\beta\sigma_1^2 F)\log\tfrac{1}{\delta} + C_2\sigma_0\sqrt{8T}.$$

Substituting $C_2$ and dividing through by $T$ gives the theorem (due to the $O$ notation in the theorem we state a $1-\delta$ probability guarantee instead of $1-3\delta$). □

### 3.2. Proof of Lemma 1

Before proving Lemma 1 we state a standard lemma (for completeness, we give a proof in Appendix A).

**Lemma 4.** *Let $g_1, \dots, g_T \in \mathbb{R}^d$ be an arbitrary sequence of vectors, and let $G_0 > 0$. For all $t \ge 1$, define*

$$G_t = \sqrt{G_0^2 + \sum_{s=1}^{t}\|g_s\|^2}.$$

*Then*

$$\sum_{t=1}^{T}\frac{\|g_t\|^2}{G_t} \le 2\sqrt{\sum_{t=1}^{T}\|g_t\|^2}, \quad \textit{and} \quad \sum_{t=1}^{T}\frac{\|g_t\|^2}{G_t^2} \le 2\log\frac{G_T}{G_0}.$$

*Proof of Lemma 1.* Using $\beta$-smoothness, we have for all $t \ge 1$ that

$$f(w_{t+1}) \le f(w_t) - \eta_t \nabla f(w_t) \cdot g_t + \tfrac{1}{2}\beta\eta_t^2\|g_t\|^2,$$

which implies (with $\Delta_t = f(w_t) - f^\star$)

$$\nabla f(w_t) \cdot g_t \le \frac{1}{\eta_t}(\Delta_t - \Delta_{t+1}) + \tfrac{1}{2}\beta\eta_t\|g_t\|^2.$$

Summing the above over $t = 1, \dots, T$, we obtain

$$\sum_{t=1}^{T}\nabla f(w_t) \cdot g_t \le \frac{\Delta_1}{\eta_0} + \sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\Delta_t + \frac{\beta}{2}\sum_{t=1}^{T}\eta_t\|g_t\|^2.$$

Now, since $\eta_t \le \eta_{t-1}$, we have for all $t$ that

$$\begin{aligned}\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &\le \eta_t\left(\frac{1}{\eta_t^2} - \frac{1}{\eta_{t-1}^2}\right) \\ &= \frac{\eta_t}{\eta^2}\left(\gamma^2 + \sum_{s=1}^{t}\|g_s\|^2 - \gamma^2 - \sum_{s=1}^{t-1}\|g_s\|^2\right) = \frac{\eta_t}{\eta^2}\|g_t\|^2,\end{aligned}$$

thus

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t \le \frac{\Delta_1}{\eta_0} + \sum_{t=1}^{T} \eta_t \|g_t\|^2 \left(\frac{\Delta_t}{\eta^2} + \frac{\beta}{2}\right).$$

Bounding $\Delta_t \le \bar{\Delta}_T$ and applying Lemma 4 on the second sum yields

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t \le \frac{\Delta_1}{\eta_0} + \left(\frac{2\bar{\Delta}_T}{\eta} + \eta\beta\right)\sqrt{\sum_{t=1}^{T} \|g_t\|^2}.$$

Plugging in the expressions for $\eta_0$ yields the lemma. □

### 3.3. Proof of Lemma 2

For our high probability analysis we use the following martingale concentration inequality (Li & Orabona, 2020, Lemma 1).

**Lemma 5.** *Assume that $Z_1, Z_2, ..., Z_T$ is a martingale difference sequence with respect to $\xi_1, \xi_2, ..., \xi_T$ and $\mathbb{E}_t\left[\exp(Z_t^2/\sigma_t^2)\right] \le \exp(1)$ for all $t$, where $\sigma_1, \ldots, \sigma_T$ is a sequence of random variables such that $\sigma_t$ is measurable with respect to $\xi_1, \xi_2, \ldots, \xi_{t-1}$. Then, for any fixed $\lambda > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that*

$$\sum_{t=1}^{T} Z_t \le \frac{3}{4}\lambda \sum_{t=1}^{T} \sigma_t^2 + \frac{1}{\lambda} \log \frac{1}{\delta}.$$

Note that it suffices to pick $\sigma_t \ge |Z_t|$ with probability 1 for the sub-Gaussian condition to hold. We also need the following technical lemmas (see Appendix A for proofs).

**Lemma 6.** *For AdaSGD we have:*

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{G_t^2} \le C_1,$$

*with $C_1$ given in Equation (3).*

**Lemma 7.** *For all $t \ge 1$,*

$$\sum_{s=1}^{t} |\tilde{\eta}_s - \eta_s|(\nabla f(w_s) \cdot g_s) \le \frac{\bar{\Delta}_t}{4} + \frac{1}{2}\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + 2\eta\sigma_0 \sum_{s=1}^{t} \frac{\|g_s\|^2}{G_s^2} + 8\eta^2\beta\sigma_1^2 \left(\sum_{s=1}^{t} \frac{\|g_s\|^2}{G_s^2}\right)^2.$$

Finally, we require a standard property of smooth functions.

**Lemma 8.** *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be $\beta$-smooth with minimum $f^\star$. Then $\|\nabla f(w)\|^2 \le 2\beta(f(w) - f^\star)$ for all $w \in \mathbb{R}^d$.*

*Proof of Lemma 2.* First we handle the randomness of $(g_t)_{t=1}^T$. Consider the sequence of RVs $(Z_t)_{t=1}^T$ defined by $Z_t = \tilde{\eta}_t \nabla f(w_t) \cdot (\nabla f(w_t) - g_t)$. Then $(Z_t)_{t=1}^T$ is a martingale difference sequence, as

$$\mathbb{E}_{t-1}[Z_t] = \tilde{\eta}_t \nabla f(w_t) \cdot (\nabla f(w_t) - \mathbb{E}_{t-1}[g_t]) = 0.$$

From Cauchy–Schwarz inequality and the noise model, $|Z_t| \le \tilde{\eta}_t \|\nabla f(w_t)\| \sqrt{\sigma_0^2 + \sigma_1^2 \|\nabla f(w_t)\|^2}$. Thus, we obtain from Lemma 5 (together with a union bound over $t$) that for any fixed $\lambda > 0$, with probability at least $1 - \delta$, for all $1 \le t \le T$,

$$\sum_{s=1}^{t} \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) \le \frac{3\lambda}{4} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|\nabla f(w_s)\|^2 (\sigma_0^2 + \sigma_1^2 \|\nabla f(w_s)\|^2) + \frac{\log \frac{T}{\delta}}{\lambda}.$$

We will proceed by induction. The claim is immediate for $t = 1$ and we move to prove it for $t + 1$. Using smoothness,

$$f(w_{s+1}) \le f(w_s) - \eta_s \nabla f(w_s) \cdot g_s + \frac{\beta}{2}\eta_s^2 \|g_s\|^2.$$

Summing the above over $s = 1, \ldots, t$,

$$f(w_{t+1}) - f(w_1) \le -\sum_{s=1}^{t} \eta_s \nabla f(w_s) \cdot g_s + \frac{\beta}{2}\sum_{s=1}^{t} \eta_s^2 \|g_s\|^2. \quad (4)$$

We move to bound the first term of the right hand side.

$$-\sum_{s=1}^{t} \eta_s \nabla f(w_s) \cdot g_s = -\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + \sum_{s=1}^{t} \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) \quad (*)$$
$$+ \sum_{s=1}^{t} (\tilde{\eta}_s - \eta_s) \nabla f(w_s) \cdot g_s. \quad (**)$$

We will bound $(*)$ and $(**)$. From Lemma 8 and our induction, $\|\nabla f(w_t)\|^2 \le 2\beta(f(w_t) - f^\star) \le 2\beta F$. Returning to our martingale bound, with the fact that $\tilde{\eta}_s \le \eta \Big/ \sqrt{\sigma_0^2 + \sigma_1^2 \|\nabla f(w_s)\|^2}$, we have

$$\sum_{s=1}^{t} \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) \le \frac{3\lambda\eta}{4} \sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 \sqrt{\sigma_0^2 + \sigma_1^2 \|\nabla f(w_s)\|^2} + \frac{\log \frac{T}{\delta}}{\lambda}$$
$$\le \frac{3\lambda\eta}{4} \sqrt{\sigma_0^2 + 2\beta\sigma_1^2 F} \sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + \frac{\log \frac{T}{\delta}}{\lambda}.$$

Setting $\lambda = 2\Big/3\eta\sqrt{\sigma_0^2 + 2\beta\sigma_1^2 F}$,

$$\sum_{s=1}^{t} \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s)$$
$$\leq \frac{1}{2}\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + \frac{3\eta}{2}\sqrt{\sigma_0^2 + 2\beta\sigma_1^2 F}\log\tfrac{T}{\delta}$$
$$\leq \frac{1}{2}\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + \frac{3\eta}{2}\sigma_0 \log\tfrac{T}{\delta} + \frac{3\eta}{2}\sqrt{2\beta\sigma_1^2 F}\log\tfrac{T}{\delta}.$$

Using the elementary fact that $ab \leq a^2/2 + b^2/2$ with $a = \sqrt{F/2}$ and $b = 3\eta\sqrt{\beta}\sigma_1 \log\frac{T}{\delta}$,

$$\sum_{s=1}^{t} \tilde{\eta}_s \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) \leq \frac{1}{2}\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2$$
$$+ \frac{3}{2}\eta\sigma_0 \log\tfrac{T}{\delta} + \frac{1}{4}F + \frac{9}{2}\eta^2\beta\sigma_1^2 \log^2\tfrac{T}{\delta}.$$

In order to bound $(**)$, using Lemma 7 with $\bar{\Delta}_t \leq F$ by induction,

$$(**) \leq \frac{F}{4} + \frac{1}{2}\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + 2\eta\sigma_0 \sum_{s=1}^{t} \frac{\|g_s\|^2}{G_s^2}$$
$$+ 8\eta^2\beta\sigma_1^2 \left(\sum_{s=1}^{t} \frac{\|g_s\|^2}{G_s^2}\right)^2$$
$$\leq \frac{F}{4} + \frac{1}{2}\sum_{s=1}^{t} \tilde{\eta}_s \|\nabla f(w_s)\|^2 + 2\eta\sigma_0 C_1 + 8\eta^2\beta\sigma_1^2 C_1^2,$$

where that last inequality follows from Lemma 6. Hence, applying the bounds $(*)$ and $(**)$,

$$-\sum_{s=1}^{t} \eta_s \nabla f(w_s) \cdot g_s \leq \tfrac{1}{2}F + \left(\tfrac{3}{2}\log\tfrac{T}{\delta} + 2C_1\right)\eta\sigma_0$$
$$+ \left(\tfrac{9}{2}\log^2\tfrac{T}{\delta} + 8C_1^2\right)\eta^2\beta\sigma_1^2.$$

Thus, returning to Equation (4) and applying Lemma 6,

$$f(w_{t+1}) - f(w_1) \leq \tfrac{1}{2}F + \left(\tfrac{3}{2}\log\tfrac{T}{\delta} + 2C_1\right)\eta\sigma_0$$
$$+ \left(\tfrac{9}{2}\log^2\tfrac{T}{\delta} + 8C_1^2\right)\eta^2\beta\sigma_1^2 + \frac{\beta}{2}\sum_{s=1}^{t} \eta_s^2 \|g_s\|^2$$
$$\leq \tfrac{1}{2}F + \left(\tfrac{3}{2}\log\tfrac{T}{\delta} + 2C_1\right)\eta\sigma_0$$
$$+ \left(\tfrac{9}{2}\log^2\tfrac{T}{\delta} + 8C_1^2\right)\eta^2\beta\sigma_1^2 + \tfrac{1}{2}\eta^2\beta C_1$$
$$= F - (f(w_1) - f^\star).$$

We conclude the induction by adding $f(w_1) - f^\star$ to both sides, obtaining $\bar{\Delta}_{T+1} = \max_{1\leq t\leq T+1} \Delta_t \leq F$. □

## 4. Convergence analysis in the convex case

In this section we give a convergence analysis for AdaSGD provided that the objective $f$ is a convex and $\beta$-smooth function over $\mathbb{R}^d$. We prove the following:

**Theorem 2.** *Assume that $f$ is convex and $\beta$-smooth. Then for AdaSGD, for any $\delta \in (0, 1/4)$, it holds with probability at least $1 - \delta$ that*

*(i)* $$f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - f^\star \leq O\left(\left(\tfrac{D^2}{\eta} + \eta\right) + D\sqrt{\log\tfrac{1}{\delta}}\right)\frac{\sigma_0}{\sqrt{T}}$$
$$+ O\left(\tfrac{\gamma d_1^2}{\eta} + \beta(1+\sigma_1^2)\left(\tfrac{D^2}{\eta} + \eta\right)^2 + \beta\sigma_1^2 D^2 \log\tfrac{1}{\delta}\right)\frac{1}{T};$$

*(ii) and $\|w_t - w^\star\| \leq D$ for all $1 \leq t \leq T+1$,*

*where $d_1 = \|w_1 - w^\star\|$, and*

$$D^2 = \widetilde{O}\Big(d_1^2 + \eta^2\big(1 + \sigma_1^2 + \gamma^{-2}(\sigma_0^2 + \beta\sigma_1^2 F)$$
$$+ \gamma^{-4}(\sigma_0^4 + \beta^2\sigma_1^4 F^2)\big)\Big),$$
$$F = \widetilde{O}\Big(f(w_1) - f^\star + \eta\sigma_0 + \eta^2\beta\sigma_1^2 + \eta^2\beta\Big).$$

Similarly to the previous section, the bound applies to the case of bounded affine noise, and we defer the extension to sub-Gaussian noise to Theorem 4 in Appendix C.

Few additional remarks are in order. First, note that when $\sigma_0 \to 0$ we obtain a rate of order $1/T$, matching the optimal rate for non-stochastic smooth optimization (note that this holds even when $\sigma_1 \gg 0$). Second, unlike the result in the non-convex case, the bound here contains a polynomial dependence on $1/\gamma$ (in the $D$ term). This is not an artifact of our analysis: setting a overly small $\gamma$ can in fact lead AdaSGD to gain a distance of $\Omega(T)$ to a minimizer; see Appendix D for more details. That said, the algorithm is fairly robust to the choice of $\gamma$; for example, consider the noisy regime with no affine noise (i.e., $\sigma_0 \gg 0$, $\sigma_1 = 0$): setting $\eta = \Theta(\|w_1 - w^\star\|)$ and any $\gamma$ within the wide range $\gamma \in [\sigma_0, \sigma_0\sqrt{T}]$ leads to the optimal rate (up to logarithmic factors).

### 4.1. Proof of Theorem 2

In the following, we will require some additional notation. First, for all $t$ we let $d_t \triangleq \|w_t - w^\star\|$ and $\bar{d}_t \triangleq \max_{s\leq t}\|w_s - w^\star\|$. In the proof below, we use a slightly different "decorrelated step size," given by

$$\hat{\eta}_s \triangleq \frac{\eta}{\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}}.$$

The proof of the main result is based on a regret bound (Lemma 10) and a bound $d_t \leq D$ (Lemma 11) which holds with high probability, analogous to Lemmas 1 and 2 of the previous section. See Appendix B for their statements and proofs. We now prove our main result in this section.

*Proof of Theorem 2.* With probability at least $1-3\delta$ we have from Lemma 11 that $d_t \le D$ for all $t \le T+1$. Using Lemma 10 we have

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \le \left(\frac{D^2}{\eta} + \eta\right)\sqrt{\sum_{t=1}^{T}\|g_t\|^2} + \frac{\gamma d_1^2}{2\eta}.$$

Under the noise assumption, smoothness and convexity,

$$\begin{aligned}\|g_t\|^2 &\le 2\|\nabla f(w_t)\|^2 + 2\|g_t - \nabla f(w_t)\|^2 \\ &\le 2(1+\sigma_1^2)\|\nabla f(w_t)\|^2 + 2\sigma_0^2 \\ &\le 4\beta(1+\sigma_1^2)\nabla f(w_t)\cdot(w_t - w^\star) + 2\sigma_0^2.\end{aligned}$$

We obtain

$$\begin{aligned}\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) &\le \left(\frac{D^2}{\eta} + \eta\right)\sigma_0\sqrt{2T} + \frac{\gamma d_1^2}{2\eta} \\ &+ 2\left(\frac{D^2}{\eta} + \eta\right)\sqrt{\beta(1+\sigma_1^2)\sum_{t=1}^{T}\nabla f(w_t)\cdot(w_t - w^\star)}.\end{aligned}$$

Using the inequality $2ab \le a^2 + b^2$, we then have

$$\begin{aligned}\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) &\le \left(\frac{D^2}{\eta} + \eta\right)\sigma_0\sqrt{2T} + \frac{\gamma d_1^2}{2\eta} \\ &+ \frac{1}{4}\sum_{t=1}^{T}\nabla f(w_t)\cdot(w_t - w^\star) + 4\beta(1+\sigma_1^2)\left(\frac{D^2}{\eta} + \eta\right)^2.\end{aligned} \tag{5}$$

By Lemma 5 with $Z_t = (\nabla f(w_t) - g_t)\cdot(w_t - w^\star)$, $|Z_t| \le \sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2}\|w_t - w^\star\|$, and

$$\lambda = \frac{1}{\sigma_0 D\sqrt{T/\log\frac{1}{\delta}} + 6\beta\sigma_1^2 D^2},$$

with probability at least $1-4\delta$ (union bound with the event of Lemma 11),

$$\begin{aligned}&\sum_{t=1}^{T}(\nabla f(w_t) - g_t)\cdot(w_t - w^\star) \\ &\le \frac{3\lambda}{4}\sum_{t=1}^{T}\left(\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2\right)\|w_t - w^\star\|^2 + \frac{1}{\lambda}\log\frac{1}{\delta} \\ &\le \sigma_0\sqrt{T\log\frac{1}{\delta}}\left(\frac{\bar{d}_T^2}{D} + D\right) + \frac{\bar{d}_T^2\sum_{t=1}^{T}\|\nabla f(w_t)\|^2}{8\beta D^2} \\ &+ 6\beta\sigma_1^2 D^2\log\frac{1}{\delta}.\end{aligned}$$

Using the bound $\bar{d}_T \le D$, with probability at least $1-4\delta$,

$$\begin{aligned}&\sum_{t=1}^{T}(\nabla f(w_t) - g_t)\cdot(w_t - w^\star) \\ &\le 2\sigma_0 D\sqrt{T\log\frac{1}{\delta}} + \frac{1}{8\beta}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 + 6\beta\sigma_1^2 D^2\log\frac{1}{\delta} \\ &\le 2\sigma_0 D\sqrt{T\log\frac{1}{\delta}} + \frac{1}{4}\sum_{t=1}^{T}\nabla f(w_t)\cdot(w_t - w^\star) \\ &+ 6\beta\sigma_1^2 D^2\log\frac{1}{\delta},\end{aligned}$$

where the last inequality follows from Lemma 8 and convexity. Summing with Equation (5) and rearranging, with probability at least $1-4\delta$,

$$\begin{aligned}\sum_{s=1}^{t}\nabla f(w_s)\cdot(w_s - w^\star) &\le 8\beta(1+\sigma_1^2)\left(\frac{D^2}{\eta} + \eta\right)^2 \\ &+ \left(\left(\frac{D^2}{\eta} + \eta\right)\sqrt{8} + 4D\sqrt{\log\frac{1}{\delta}}\right)\sigma_0\sqrt{T} + \frac{\gamma d_1^2}{\eta} \\ &+ 12\beta\sigma_1^2 D^2\log\frac{1}{\delta}.\end{aligned}$$

A standard application of Jensen's inequality concludes the proof (up to a constant factor in the failure probability, which we subsumed in the big-O notation). □

## Acknowledgements

This work has received funding from the Israel Science Foundation (ISF, grant number 993/17), Len Blavatnik and the Blavatnik Family foundation, the Adelis Research Fund for Artificial Intelligence, the Yandex Initiative for Machine Learning at Tel Aviv University, and from the Tel Aviv University Center for AI and Data Science (TAD).

## References

Alacaoglu, A., Malitsky, Y., Mertikopoulos, P., and Cevher, V. A new regret analysis for adam-type algorithms. In *International conference on machine learning*, pp. 202–210. PMLR, 2020.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50, 2022.

Bottou, L. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pp. 421–436. Springer, 2012.

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

Carmon, Y. and Hinder, O. Making SGD parameter-free. In *COLT*, 2022.

Cutkosky, A. and Mehta, H. High-probability bounds for non-convex stochastic optimization with heavy tails. In *NeurIPS*, 2021.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

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

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. A. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *COLT*, 2022.

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

Harvey, N. J., Liaw, C., Plan, Y., and Randhawa, S. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

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

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

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

Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992. PMLR, 2019.

Li, X. and Orabona, F. A high probability analysis of adaptive sgd with momentum. In *Workshop on Beyond First Order Methods in ML Systems at ICML'20*, 2020.

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. High probability convergence of stochastic gradient methods. *arXiv preprint arXiv:2302.14843*, 2023.

Mhammedi, Z. and Koolen, W. M. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, pp. 2858–2887. PMLR, 2020.

Orabona, F. and Pál, D. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

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

Schaul, T., Zhang, S., and LeCun, Y. No more pesky learning rates. In *International conference on machine learning*, pp. 343–351. PMLR, 2013.

Tran, P. T. et al. On the convergence proof of amsgrad and a new version. *IEEE Access*, 7:61706–61716, 2019.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686. PMLR, 2019.

Zhou, D., Chen, J., Cao, Y., Tang, Y., Yang, Z., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

# A. Proofs of Section 3

## A.1. Proof of Lemma 4

*Proof.* Define $\tilde{G}_t = \sqrt{G_t^2 - G_0^2}$. Observe that for any $t \geq 1$,

$$\|g_t\|^2 = \sum_{s=1}^{t} \|g_s\|^2 - \sum_{s=1}^{t-1} \|g_s\|^2 = \tilde{G}_t^2 - \tilde{G}_{t-1}^2 = \big(\tilde{G}_t + \tilde{G}_{t-1}\big)\big(\tilde{G}_t - \tilde{G}_{t-1}\big).$$

Since $\tilde{G}_t \geq \tilde{G}_{t-1}$, this implies

$$\frac{\|g_t\|^2}{\tilde{G}_t} = \left(\frac{\tilde{G}_t + \tilde{G}_{t-1}}{\tilde{G}_t}\right)\big(\tilde{G}_t - \tilde{G}_{t-1}\big) \leq 2\big(\tilde{G}_t - \tilde{G}_{t-1}\big).$$

Summing this over $t = 1, \ldots, T$ and telescoping terms, we obtain

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{G_t} \leq \sum_{t=1}^{T} \frac{\|g_t\|^2}{\tilde{G}_t} \leq 2\tilde{G}_T = 2\sqrt{\sum_{t=1}^{T} \|g_s\|^2},$$

where the first inequality follows by $G_t \geq \tilde{G}_t$. On the other hand, using the inequality $1 - x \leq \log(1/x)$,

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{G_t^2} = \sum_{t=1}^{T} \left(1 - \frac{G_{t-1}^2}{G_t^2}\right) \leq \sum_{t=1}^{T} \log \frac{G_t^2}{G_{t-1}^2} = 2 \log \frac{G_T}{G_0}.$$ □

## A.2. Proof of Lemma 8

*Proof.* Let $w^+ = w - \frac{1}{\beta}\nabla f(w)$. Then from smoothness,

$$f(w^+) - f(w) \leq \nabla f(w) \cdot (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 = -\frac{1}{2\beta}\|\nabla f(w)\|^2.$$

Hence,

$$\|\nabla f(w)\|^2 \leq 2\beta(f(w) - f(w^+)) \leq 2\beta(f(w) - f^\star).$$ □

## A.3. Proof of Lemma 6

*Proof.* Using Lemma 4,

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{G_t^2} \leq 2 \log \frac{G_T}{G_0}.$$

Since $\|g_t\|^2 \leq 2\|\nabla f(w_t)\|^2 + 2\|g_t - \nabla f(w_t)\|^2 \leq 2(1 + \sigma_1^2)\|\nabla f(w_t)\|^2 + 2\sigma_0^2$,

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{G_t^2} \leq \log\left(G_0^2 + 2\sigma_0^2 T + 2(1 + \sigma_1^2)\sum_{t=1}^{T} \|\nabla f(w_t)\|^2\right) + 2\log(1/G_0).$$

From smoothness,

$$\begin{aligned}\|\nabla f(w_t)\| &\leq \|\nabla f(w_{t-1})\| + \|\nabla f(w_t) - \nabla f(w_{t-1})\| \leq \|\nabla f(w_{t-1})\| + \beta\|w_t - w_{t-1}\| \\ &= \|\nabla f(w_{t-1})\| + \beta\eta_t\|g_t\| \leq \|\nabla f(w_{t-1})\| + \eta\beta.\end{aligned}$$

Thus,

$$\|\nabla f(w_t)\| \leq \|\nabla f(w_1)\| + \eta\beta(t-1) \leq \|\nabla f(w_1)\| + \eta\beta t.$$

summing from $t = 1, \dots, T$,

$$\begin{aligned}
\sum_{t=1}^{T} \|\nabla f(w_t)\|^2 &\le 2\sum_{t=1}^{T}(\eta^2\beta^2 t^2 + \|\nabla f(w_1)\|^2) && (\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2)\\
&\le 2\eta^2\beta^2 T^3 + 2\|\nabla f(w_1)\|^2 T && (\textstyle\sum_{t=1}^{T} t^2 \le T^3)\\
&\le 2\eta^2\beta^2 T^3 + 4\beta(f(w_1) - f^\star)T. && \text{(from smoothness, Lemma 8)}
\end{aligned}$$

Hence, with $G_0^2 = \gamma^2$,

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{G_t^2} \le \log\left(1 + \frac{2\sigma_0^2 T + 4(1+\sigma_1^2)\eta^2\beta^2 T^3 + 8(1+\sigma_1^2)\beta(f(w_1) - f^\star)T}{\gamma^2}\right) \le C_1. \qquad \square$$

### A.4. Proof of Lemma 7

In order to prove the bound we also need the following technical lemma (proof follows).

**Lemma 9.** *For all* $1 \le s \le T$,

$$|\tilde{\eta}_s - \eta_s| \le \frac{2\tilde{\eta}_s\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}}{G_s}.$$

*Proof of Lemma 7.* By Lemma 9,

$$|\tilde{\eta}_s - \eta_s| \le \frac{2\tilde{\eta}_s\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}}{G_s}.$$

Hence, by applying Cauchy-Schwarz inequality and using $ab \le \frac{a^2}{2} + \frac{b^2}{2}$ with $a = \sqrt{\tilde{\eta}_s}\|\nabla f(w_s)\|$ and $b = 2\sqrt{\tilde{\eta}_s(\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2)}\frac{\|g_s\|}{G_s}$,

$$\begin{aligned}
|\tilde{\eta}_s - \eta_s|\nabla f(w_s)\cdot g_s &\le 2\tilde{\eta}_s\|\nabla f(w_s)\|\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}\frac{\|g_s\|}{G_s}\\
&\le \frac{1}{2}\tilde{\eta}_s\|\nabla f(w_s)\|^2 + 2\tilde{\eta}_s\left(\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2\right)\frac{\|g_s\|^2}{G_s^2}\\
&\le \frac{1}{2}\tilde{\eta}_s\|\nabla f(w_s)\|^2 + 2\eta\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}\frac{\|g_s\|^2}{G_s^2}. && (\tilde{\eta}_s \le \eta/\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2})
\end{aligned}$$

Using the smoothness of $f$, $\|\nabla f(w_s)\|^2 \le 2\beta(f(w_s) - f^\star) \le 2\beta\bar{\Delta}_t$ (Lemma 8). Thus, summing for $s = 1, \dots, t$,

$$\sum_{s=1}^{t}|\tilde{\eta}_s - \eta_s|\nabla f(w_s)\cdot g_s \le \frac{1}{2}\sum_{s=1}^{t}\tilde{\eta}_s\|\nabla f(w_s)\|^2 + 2\eta\sqrt{\sigma_0^2 + 2\beta\sigma_1^2\bar{\Delta}_t}\sum_{s=1}^{t}\frac{\|g_s\|^2}{G_s^2}.$$

Again, using $ab \le \frac{a^2}{2} + \frac{b^2}{2}$ with $a = \sqrt{\bar{\Delta}_t/2}$ and $b = 4\eta\sigma_1\sqrt{\beta}\sum_{s=1}^{t}\frac{\|g_s\|^2}{G_s^2}$,

$$\begin{aligned}
\sum_{s=1}^{t}|\tilde{\eta}_s - \eta_s|\nabla f(w_s)\cdot g_s &\le \frac{1}{2}\sum_{s=1}^{t}\tilde{\eta}_s\|\nabla f(w_s)\|^2 + 2\eta\sigma_0\sum_{s=1}^{t}\frac{\|g_s\|^2}{G_s^2} + 2\eta\sqrt{2\beta\sigma_1^2\bar{\Delta}_t}\sum_{s=1}^{t}\frac{\|g_s\|^2}{G_s^2}\\
&\le \frac{1}{2}\sum_{s=1}^{t}\tilde{\eta}_s\|\nabla f(w_s)\|^2 + 2\eta\sigma_0\sum_{s=1}^{t}\frac{\|g_s\|^2}{G_s^2} + \frac{\bar{\Delta}_t}{4} + 8\eta^2\beta\sigma_1^2\left(\sum_{s=1}^{t}\frac{\|g_s\|^2}{G_s^2}\right)^2. \qquad \square
\end{aligned}$$

*Proof of Lemma 9.* Using $\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}} = \frac{b-a}{\sqrt{a}\sqrt{b}(\sqrt{a}+\sqrt{b})}$,

$$\begin{aligned}
\tilde{\eta}_s - \eta_s &= \frac{\eta}{\sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2}} - \frac{\eta}{\sqrt{G_{s-1}^2 + \|g_s\|^2}} \\
&= \frac{\eta(\|g_s\|^2 - \sigma_0^2 - (1+\sigma_1^2)\|\nabla f(w_s)\|^2)}{G_s\sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2}(G_s + \sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2})} \\
&= \frac{\tilde{\eta}_s}{G_s} \cdot \frac{\|g_s\|^2 - \sigma_0^2 - (1+\sigma_1^2)\|\nabla f(w_s)\|^2}{G_s + \sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
|\tilde{\eta}_s - \eta_s| &= \frac{\tilde{\eta}_s}{G_s} \cdot \frac{\left|\|g_s\|^2 - \sigma_0^2 - (1+\sigma_1^2)\|\nabla f(w_s)\|^2\right|}{G_s + \sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2}} \\
&\le \frac{\tilde{\eta}_s}{G_s} \cdot \frac{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2 + \|g_s - \nabla f(w_s)\|\|g_s + \nabla f(w_s)\|}{G_s + \sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2}}.
\end{aligned}$$

Since $\|g_s + \nabla f(w_s)\| \le \|g_s\| + \|\nabla f(w_s)\| \le G_s + \sqrt{G_{s-1}^2 + \sigma_0^2 + (1+\sigma_1^2)\|\nabla f(w_s)\|^2}$,

$$|\tilde{\eta}_s - \eta_s| \le \frac{\tilde{\eta}_s}{G_s} \cdot \left(\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2} + \|g_s - \nabla f(w_s)\|\right) \le \frac{2\tilde{\eta}_s\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}}{G_s}. \qquad \square$$

### A.5. Proof of Lemma 3

*Proof of Lemma 3.* Consider the sequence of R.V.s $(Z_t)_{t=1}^T$ defined by $Z_t = \nabla f(w_t) \cdot (\nabla f(w_t) - g_t)$. Then $(Z_t)_{t=1}^T$ is a martingale difference sequence, as

$$\mathbb{E}_{t-1}[Z_t] = \nabla f(w_t) \cdot (\nabla f(w_t) - \mathbb{E}_{t-1}[g_t]) = 0.$$

By Cauchy-Schwarz inequality and the noise assumption,

$$|Z_t| \le \|\nabla f(w_t)\|\|\nabla f(w_t) - g_t\| \le \|\nabla f(w_t)\|\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2}.$$

Thus, we obtain from Lemma 5 with

$$|Z_t| \le \|\nabla f(w_t)\|\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2}$$

and $\lambda = 1/3(\sigma_0^2 + 2\beta\sigma_1^2 F)$ that with probability at least $1-\delta$,

$$\sum_{t=1}^T \nabla f(w_t) \cdot (\nabla f(w_t) - g_t) \le \frac{\sum_{t=1}^T(\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2)\|\nabla f(w_t)\|^2}{4(\sigma_0^2 + 2\beta\sigma_1^2 F)} + 3(\sigma_0^2 + 2\beta\sigma_1^2 F)\log(1/\delta).$$

From Lemmas 2 and 8, with probability at least $1-\delta$, $\|\nabla f(w_t)\|^2 \le 2\beta F$. Hence, under a union bound with Lemma 2, with probability at least $1-2\delta$,

$$\sum_{t=1}^T \nabla f(w_t) \cdot (\nabla f(w_t) - g_t) \le \frac{1}{4}\sum_{t=1}^T\|\nabla f(w_t)\|^2 + 3(\sigma_0^2 + 2\beta\sigma_1^2 F)\log(1/\delta). \qquad \square$$

## B. Proofs of Section 4

Below are the statements of the two main lemmas used to prove Theorem 2 (their proofs follow). The first is a regret analysis which depends on $\bar{d}_t$.

**Lemma 10.** *Let $f$ be a convex and $\beta$-smooth function. Then*

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \le \left(\frac{\bar{d}_T^2}{\eta} + \eta\right)\sqrt{\sum_{t=1}^{T}\|g_t\|^2} + \frac{\gamma d_1^2}{2\eta}.$$

Similarly to the non-convex case, we need a high probability bound of $\|w_t - w^\star\|$ in order to use the regret analysis of Lemma 10; this is given in our next lemma.

**Lemma 11.** *With probability at least $1 - 3\delta$, it holds that for all $t \le T + 1$,*

$$d_t^2 \le D^2 \triangleq 2d_1^2 + \eta^2\left(\frac{1}{2} + 2C_1 + 8C^2 + A_T C + B_T\left(\frac{\sigma_0}{\gamma} + \sigma_1\right)^2\right),$$

*where $C = 2\log(1 + \sigma^2 T/2\gamma^2) + 7\sigma^2\log(T/\delta)/\gamma^2$, $C_1$ is defined at Lemma 6, $A_T = 512\log\left(\frac{60\log^2(6T)}{\delta}\right)$ and $B_T = 512\log^2\left(\frac{60\log^2(6T)}{\delta}\right)$.*

### B.1. Proof of Lemma 10

*Proof.* From the standard analysis of SGD, we have for all $t$ that

$$g_t \cdot (w_t - w^\star) = \frac{1}{2\eta_t}\left(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\right) + \frac{\eta_t}{2}\|g_t\|^2.$$

Summing over $t = 1, \dots, T$, we have

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \le \frac{\gamma\|w_1 - w^\star\|^2}{2\eta} + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\|w_t - w^\star\|^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|^2.$$

Now, since $\eta_t \le \eta_{t-1}$, we have for all $t$ that

$$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \le \eta_t\left(\frac{1}{\eta_t^2} - \frac{1}{\eta_{t-1}^2}\right) = \frac{\eta_t}{\eta^2}\left(\gamma^2 + \sum_{s=1}^{t}\|g_s\|^2 - \gamma^2 - \sum_{s=1}^{t-1}\|g_s\|^2\right) = \frac{\eta_t}{\eta^2}\|g_t\|^2,$$

thus

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \le \frac{\gamma\|w_1 - w^\star\|^2}{2\eta} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|^2\left(1 + \frac{\|w_t - w^\star\|^2}{\eta^2}\right).$$

Finally, bounding $\|w_t - w^\star\| \le \bar{d}_T$ and applying Lemma 4 we obtain

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \le \left(\frac{\bar{d}_T^2}{\eta} + \eta\right)\sqrt{\sum_{t=1}^{T}\|g_t\|^2} + \frac{\gamma\|w_1 - w^\star\|^2}{2\eta}.$$ □

### B.2. Proof of Lemma 11

In order to obtain the high probability bound of $\bar{d}_T$ we use of the following empirical Bernstein concentration bound.

**Lemma 12** (Corollary 1 of Carmon & Hinder (2022))**.** *Let $X_t$ be adapted to $\mathcal{F}_t$ such that $|X_t| \le 1$ with probability 1 for all $t$. Then, for every $\delta \in (0, 1)$ and any $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le 1$ with probability 1,*

$$\Pr\left(\exists t < \infty : \left|\sum_{s \le t}(X_s - \mathbb{E}[X_s|\mathcal{F}_{s-1}])\right| \ge \sqrt{A_t(\delta)\sum_{s \le t}(X_s - \hat{X}_s)^2 + B_t(\delta)}\right) \le \delta,$$

*where $A_t(\delta) = 16\log\left(\frac{60\log(6t)}{\delta}\right)$ and $B_t(\delta) = 16\log^2\left(\frac{60\log(6t)}{\delta}\right)$.*

We use the concentration bound to prove the following technical martingale bound.

**Lemma 13.** *Let $\bar{d}'_t = \max\{\bar{d}_t, \eta\}$. Then with probability at least $1-\delta$, it holds that for all $t \le T$,*

$$\sum_{s=1}^{t} \hat{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w^\star) \le 2\bar{d}'_t \sqrt{A_t(\delta/\log(4T)) \sum_{s \le t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 + \eta^2 \left(\frac{\sigma_0}{\gamma} + \sigma_1\right)^2 B_t(\delta/\log(4T))},$$

*where $A(\cdot)$ and $B(\cdot)$ are defined at Lemma 12.*

The next lemma bounds the $\sum_{s \le t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2$ term which appear in Lemma 13.

**Lemma 14.** *With probability at least $1-2\delta$, for all $1 \le t \le T$,*

$$\sum_{s=1}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 \le 2\eta^2 \log(1 + \sigma^2 T/2\gamma^2) + 7\eta_0^2 \sigma^2 \log(T/\delta).$$

Now we move to prove the lemma.

*Proof of Lemma 11.* Rolling a single step of SGD,

$$\|w_{s+1} - w^\star\|^2 = \|w_s - w^\star\|^2 - 2\eta_s g_s \cdot (w_s - w^\star) + \eta_s^2 \|g_s\|^2.$$

Summing for $s = 1 \dots t$ and applying Lemma 6,

$$\begin{aligned} \|w_{t+1} - w^\star\|^2 &= \|w_1 - w^\star\|^2 - 2\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w^\star) + \sum_{s=1}^{t} \eta_s^2 \|g_s\|^2 \\ &\le \|w_1 - w^\star\|^2 \underbrace{-2\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w^\star)}_{(*)} + \eta^2 C_1. \end{aligned} \tag{6}$$

Focusing on $(*)$, due to convexity,

$$\begin{aligned} -2\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w^\star) &= -2\sum_{s=1}^{t} \eta_s \nabla f(w_s) \cdot (w_s - w^\star) + 2\sum_{s=1}^{t} \eta_s (\nabla f(w_s) - g_s) \cdot (w_s - w^\star) \\ &\le 2\sum_{s=1}^{t} \eta_s (\nabla f(w_s) - g_s) \cdot (w_s - w^\star). \end{aligned}$$

In order to create a martingale we replace $\eta_s = \eta/\sqrt{G_{s-1}^2 + \|g_s\|^2}$ with $\hat{\eta}_s = \eta/\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}$,

$$(*) \le 2\sum_{s=1}^{t} \hat{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w^\star) + 2\sum_{s=1}^{t} (\eta_s - \hat{\eta}_s)(\nabla f(w_s) - g_s) \cdot (w_s - w^\star).$$

We observe that

$$
\begin{aligned}
|\eta_s - \hat{\eta}_s| &= \eta \frac{\left|\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2} - \sqrt{G_{s-1}^2 + \|g_s\|^2}\right|}{\sqrt{G_{s-1}^2 + \|g_s\|^2}\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}} \\
&= \eta \frac{\left|\|\nabla f(w_s)\|^2 - \|g_s\|^2\right|}{\sqrt{G_{s-1}^2 + \|g_s\|^2}\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}(\sqrt{G_{s-1}^2 + \|g_s\|^2} + \sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2})} \\
&\le \eta \frac{\|\nabla f(w_s) - g_s\|(\|\nabla f(w_s)\| + \|g_s\|)}{\sqrt{G_{s-1}^2 + \|g_s\|^2}\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}(\sqrt{G_{s-1}^2 + \|g_s\|^2} + \sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2})} \\
&\le \eta \frac{\|\nabla f(w_s) - g_s\|}{\sqrt{G_{s-1}^2 + \|g_s\|^2}\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\sum_{s=1}^{t} (\eta_s - \hat{\eta}_s)(\nabla f(w_s) - g_s) \cdot (w_s - w^\star) &\le \sum_{s=1}^{t} |\eta_s - \hat{\eta}_s| \|\nabla f(w_s) - g_s\| d_s \\
&\le \bar{d}_t \sum_{s=1}^{t} |\eta_s - \hat{\eta}_s| \|\nabla f(w_s) - g_s\| \\
&\le \eta \bar{d}_t \sum_{s=1}^{t} \frac{\|\nabla f(w_s) - g_s\|^2}{\sqrt{G_{s-1}^2 + \|g_s\|^2}\sqrt{G_{s-1}^2 + \|\nabla f(w_s)\|^2}} \\
&\le \frac{\bar{d}_t}{\eta} \sum_{s=1}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2.
\end{aligned}
$$

Combining the inequality with Lemma 13, with probability at least $1 - \delta$,

$$
\begin{aligned}
(*) \le\ & \frac{2\bar{d}_t}{\eta} \sum_{s=1}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 \\
& + 4\bar{d'}_t \sqrt{A_t(\delta/\log(4T)) \sum_{s \le t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 + \eta^2 \left(\frac{\sigma_0}{\gamma} + \sigma_1\right)^2 B_t(\delta/\log(4T))},
\end{aligned}
$$

and using $ab \le a^2/2 + b^2/2$,

$$
\begin{aligned}
(*) \le\ & \frac{\bar{d}_t^2}{4} + \frac{4}{\eta^2}\left(\sum_{s=1}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2\right)^2 + \frac{\bar{d'}_t^2}{4} \\
& + 16\left(A_t(\delta/\log(4T)) \sum_{s \le t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 + \eta^2 \left(\frac{\sigma_0}{\gamma} + \sigma_1\right)^2 B_t(\delta/\log(4T))\right) \\
\le\ & \frac{\bar{d}_t^2}{2} + \frac{4}{\eta^2}\left(\sum_{s=1}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2\right)^2 + \frac{\eta^2}{4} \qquad (\bar{d'}_t = \max\{\bar{d}_t, \eta\}) \\
& + 16\left(A_t(\delta/\log(4T)) \sum_{s \le t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 + \eta^2 \left(\frac{\sigma_0}{\gamma} + \sigma_1\right)^2 B_t(\delta/\log(4T))\right).
\end{aligned}
$$

Applying Lemma 14, with probability at least $1 - 3\delta$ (via a union bound),

$$
(*) \le \frac{\bar{d}_t^2}{2} + 4\eta^2 C^2 + \frac{\eta^2}{4} + 16\eta^2 \left(A_t(\delta/\log(4T))C + \left(\frac{\sigma_0}{\gamma} + \sigma_1\right)^2 B_t(\delta/\log(4T))\right),
$$

where $C = 2\log(1+\sigma^2 T/2\gamma^2) + 7\sigma^2/\gamma^2 \log(T/\delta)$. Hence, returning to Equation (6) and applying Lemma 6,

$$\begin{aligned} d_{t+1}^2 &\le \frac{\bar{d}_t^2}{2} + d_1^2 + \eta^2\left(\frac{1}{4} + C_1 + 4C^2 + 16\left(A_t(\delta/\log(4T))C + \left(\frac{\sigma_0}{\gamma}+\sigma_1\right)^2 B_t(\delta/\log(4T))\right)\right) \\ &\le \frac{\bar{d}_t^2}{2} + \frac{D}{2}. \qquad \text{(def of } D,\ 32A_t(\delta/\log(4T)) \le A_T \text{ and } 32B_t(\delta/\log(4T)) \le B_T) \end{aligned}$$

Hence, rolling to $t = 1$, with probability at least $1 - 3\delta$, $d_t \le D$ for all $1 \le t \le T+1$. □

*Proof of Lemma 13.* In order to invoke Lemma 12 we will replace $w_s - w^\star$ with a version which is scaled and projected to the unit ball. We denote $a_k = 2^{k-1}\bar{d}'_1$ and $k_t = \lceil \log(\bar{d}'_t/\bar{d}'_1)\rceil + 1$. Thus, $\bar{d}_t \le \bar{d}'_t \le a_{k_t} \le 2\bar{d}'_t$. Since $\|w_{s+1} - w^\star\| \le \|w_s - w^\star\| + \eta$ for all $s$, $\bar{d}_t \le d_1 + \eta(t-1)$ and $1 \le k_t \le \lceil \log(t)\rceil + 1 \le \log(4T)$. Defining the projection to the unit ball, $\Pi_1(x) = x/\max\{1, \|x\|\}$,

$$\frac{w_s - w^\star}{a_{k_t}} = \Pi_1\left(\frac{w_s - w^\star}{a_{k_t}}\right).$$

To simplify notation, let $C = \eta(\sigma_0/\gamma + \sigma_1)$ and note that

$$\|\hat{\eta}_s(\nabla f(w_s) - g_s)\| \le \frac{\eta\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}}{\sqrt{\gamma^2 + \|\nabla f(w_s)\|^2}} \le \frac{\eta\sigma_0}{\gamma} + \eta\sigma_1 = C.$$

Thus,

$$\begin{aligned} &\sum_{s=1}^{t} \frac{\hat{\eta}_s(\nabla f(w_s) - g_s)\cdot(w_s - w^\star)}{C a_{k_t}} \\ &\quad= \sum_{s=1}^{t} \frac{\hat{\eta}_s(\nabla f(w_s) - g_s)}{C}\cdot \Pi_1\left(\frac{w_s - w^\star}{a_{k_t}}\right) \\ &\quad\le \left|\sum_{s=1}^{t} \frac{\hat{\eta}_s(\nabla f(w_s) - g_s)}{C}\cdot \Pi_1\left(\frac{w_s - w^\star}{a_{k_t}}\right)\right| \\ &\quad\le \max_{1\le k\le \lfloor \log(4T)\rfloor}\left|\sum_{s=1}^{t} \frac{\hat{\eta}_s(\nabla f(w_s) - g_s)}{C}\cdot \Pi_1\left(\frac{w_s - w^\star}{a_{k}}\right)\right|. \end{aligned} \tag{7}$$

Let $X_s^{(k)} = \frac{\hat{\eta}_s(\nabla f(w_s) - g_s)}{C}\cdot \Pi_1\left(\frac{w_s - w^\star}{a_k}\right)$ for some $k$.

$$\mathbb{E}_{s-1}[g_s] = \nabla f(w_s) \quad\Longrightarrow\quad \mathbb{E}_{s-1}\left[X_s^{(k)}\right] = 0.$$

Also note that $X_s^{(k)} \le 1$ with probability 1. Using Lemma 12 with the $X_s^{(k)}$ we defined and $\hat{X}_s = 0$, for any $k$ and $\delta' \in (0,1)$, with probability at least $1-\delta'$, for all $t \le T$,

$$\left|\sum_{s\le t} X_s^{(k)}\right| \le \sqrt{A_t(\delta')\sum_{s\le t}(X_s^{(k)})^2 + B_t(\delta')}.$$

We can upper bound $(X_s^{(k)})^2$,

$$\begin{aligned} (X_s^{(k)})^2 &\le \frac{\hat{\eta}_s^2\|\nabla f(w_s) - g_s\|^2}{C^2}\left\|\Pi_1\left(\frac{w_s - w^\star}{a_k}\right)\right\|^2 && \text{(Cauchy-Schwarz inequality)} \\ &\le \frac{\hat{\eta}_s^2\|\nabla f(w_s) - g_s\|^2}{C^2} && (\|\Pi_1(x)\| \le 1) \\ &\le \frac{\eta_{s-1}^2}{C^2}\|\nabla f(w_s) - g_s\|^2. && (\eta_{s-1} \ge \hat{\eta}_s) \end{aligned}$$

Thus, returning to Equation (7) multiplied by $Ca_{k_t}$, with probability at least $1-\delta'\log(4T)$ (union bound for all $1 \le k \le \lfloor \log(4T) \rfloor$),

$$\sum_{s=1}^{t} \hat{\eta}_s(\nabla f(w_s) - g_s)\cdot(w_s - w^\star) \le Ca_{k_t}\sqrt{A_t(\delta')\sum_{s\le t}\frac{\eta_{s-1}^2}{C^2}\|\nabla f(w_s)-g_s\|^2 + B_t(\delta')}.$$

As $a_{k_t} \le 2\bar{d}'_t$, picking $\delta' = \delta/\log(4T)$, with probability at least $1-\delta$,

$$\sum_{s=1}^{t} \eta_{s-1}(\nabla f(w_s) - g_s)\cdot(w_s - w^\star) \le 2\bar{d}'_t\sqrt{A_t(\delta/\log(4T))\sum_{s\le t}\eta_{s-1}^2\|\nabla f(w_s)-g_s\|^2 + C^2B_t(\delta/\log(4T))}. \qquad \square$$

*Proof of Lemma 14.* Let $Z_s = \nabla f(w_s)\cdot(\nabla f(w_s) - g_s)$ and note that $\mathbb{E}_{s-1}[Z_s] = 0$. Hence, $(Z_s)_{s=1}^t$ is a martingale difference sequence. By Lemma 5 (with $|Z_s| \le \sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2}\|\nabla f(w_s)\|$ and $\lambda = 2/3\sigma^2$) and a union bound (over $1 \le t \le T$), with probability at least $1-\delta$,

$$\begin{aligned}\sum_{s=1}^{t}\nabla f(w_s)\cdot(\nabla f(w_s)-g_s) &\le \frac{3\lambda}{4}\sum_{s=1}^{t}(\sigma_0^2+\sigma_1^2\|\nabla f(w_s)\|^2)\|\nabla f(w_s)\|^2 + \frac{1}{\lambda}\log(T/\delta)\\ &\le \frac{\sigma_0^2 + 2\beta\sigma_1^2\bar{\Delta}_t}{2\sigma^2}\sum_{s=1}^{t}\|\nabla f(w_s)\|^2 + \frac{3}{2}\sigma^2\log(T/\delta) \qquad \text{(Lemma 8)}\end{aligned}$$

for all $t \in [T]$, and under a union bound with Lemma 2, with probability at least $1-2\delta$, since $\sigma_0^2 + 2\beta\sigma_1^2\bar{\Delta}_t \le \sigma^2$,

$$\sum_{s=1}^{t}\nabla f(w_s)\cdot(\nabla f(w_s)-g_s) \le \frac{1}{2}\sum_{s=1}^{t}\|\nabla f(w_s)\|^2 + \frac{3}{2}\sigma^2\log(T/\delta). \tag{8}$$

Let

$$k = \min\left\{t \le T : \sum_{s=1}^{t}\|\nabla f(w_s)-g_s\|^2 > 6\sigma^2\log(T/\delta)\right\}.$$

For $t < k$,

$$\sum_{s=1}^{t}\eta_{s-1}^2\|\nabla f(w_s)-g_t\|^2 \le \eta_0^2\sum_{s=1}^{t}\|\nabla f(w_s)-g_t\|^2 \le 6\eta_0^2\sigma^2\log(T/\delta).$$

For $t \ge k$,

$$\begin{aligned}G_t^2 &= \gamma^2 + \sum_{s=1}^{t}\|\nabla f(w_s)\|^2 + \sum_{s=1}^{t}\|\nabla f(w_s)-g_s\|^2 - 2\sum_{s=1}^{t}\nabla f(w_s)\cdot(\nabla f(w_s)-g_s)\\ &\ge \gamma^2 + \sum_{s=1}^{t}\|\nabla f(w_s)-g_s\|^2 - 3\sigma^2\log(T/\delta) \qquad \text{(Equation (8))}\\ &\ge \gamma^2 + \frac{1}{2}\sum_{s=1}^{t}\|\nabla f(w_s)-g_s\|^2. \qquad (t \ge k)\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{s=k}^{t}\eta_{s-1}^2\|\nabla f(w_s)-g_s\|^2 &= \sum_{s=k}^{t}\eta_s^2\|\nabla f(w_s)-g_s\|^2 + \sum_{s=k}^{t}(\eta_{s-1}^2-\eta_s^2)\|\nabla f(w_s)-g_s\|^2\\ &\le \sum_{s=k}^{t}\eta_s^2\|\nabla f(w_s)-g_s\|^2 + \eta_0^2\sigma^2\\ &\le 2\eta^2\sum_{s=k}^{t}\frac{\|\nabla f(w_s)-g_s\|^2}{2\gamma^2+\sum_{k=1}^{s}\|\nabla f(w_k)-g_k\|^2} + \eta_0^2\sigma^2.\end{aligned}$$

Using Lemma 4 with $G_0 = \sqrt{2\gamma^2}$ and $g_t = \nabla f(w_s) - g_s$,

$$\begin{aligned}\sum_{s=k}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 &\le 2\eta^2 \log\left(\frac{2\gamma^2 + \sum_{s=k}^{t} \|\nabla f(w_s) - g_s\|^2}{2\gamma^2}\right) + \eta_0^2\sigma^2 \\ &\le 2\eta^2 \log\left(1 + \frac{\sigma_0^2 T + \sigma_1^2 \sum_{s=k}^{t} \|\nabla f(w_s)\|^2}{2\gamma^2}\right) + \eta_0^2\sigma^2 && \text{(noise bound)} \\ &\le 2\eta^2 \log(1 + \sigma^2 T/2\gamma^2) + \eta_0^2\sigma^2. && \text{(Lemmas 2 and 8)}\end{aligned}$$

Hence, combining with the case of $t < k$, with probability at least $1 - 2\delta$, for all $t \le T$,

$$\sum_{s=1}^{t} \eta_{s-1}^2 \|\nabla f(w_s) - g_s\|^2 \le 2\eta^2 \log(1 + \sigma^2 T/2\gamma^2) + 7\eta_0^2\sigma^2 \log(T/\delta). \qquad \square$$

## C. A General Reduction from sub-Gaussian to Bounded Noise

Given an unbiased gradient oracle one can consider different noise assumptions. For establishing high probability convergence either sub-Gaussian noise or bounded noise are most commonly used. Sub-Gaussian noise is less restrictive but can make the analysis cumbersome. In this section we explain how obtaining guarantees with a bounded oracle can yield the same guarantees with a sub-Gaussian oracle (up to logarithmic factors) with high probability.

Let $X$ be a zero-mean random vector such that for some $\sigma > 0$, for all $t > 0$,

$$\Pr(\|X\| \ge t) \le 2\exp(-t^2/\sigma^2). \tag{9}$$

This definition is equivalent to the sub-Gaussian definition used by Li & Orabona (2020) and is similar to the one used by Kavis et al. (2022).

The following lemma is a reduction from $X$ to a new bounded random vector with zero-mean which is equal to $X$ with high probability.

**Lemma 15.** *For any $\delta \in (0, 1)$ there exist a random variable $\bar{X}$ such that:*

*(i) $\bar{X}$ is zero-mean: $\mathbb{E}[\bar{X}] = 0$;*
*(ii) $\bar{X}$ is equal to $X$ w.h.p.: $\Pr(\bar{X} = X) \ge 1 - \delta$;*
*(iii) $\bar{X}$ is bounded with probability 1: $\Pr(\|\bar{X}\| < 3\sigma\sqrt{\ln(4/\delta)}) = 1$.*

The lemma implies that given a sequence of oracle queries $(w_t)_{t=1}^T$ to a $\sigma$-sub-Gaussian unbiased oracle $g(w)$ (where $w_t$ might depend on the previous queries) there exist a $O(\sigma\sqrt{\log(T/\delta')})$-bounded unbiased oracle $\tilde{g}(w)$ such that with probability at least $1 - \delta'$ return the same observations. This happens by the following procedure (Given a query $w$):

1. Let $X = g(w) - \mathbb{E}[g(w)]$.
2. Let $\bar{X}$ be the truncated random vector of Lemma 15 for $\delta = \delta'/T$.
3. Return $\mathbb{E}[g(w)] + \bar{X}$.

Hence, up to logarithmic factors, one can assume a bounded unbiased oracle and obtain the same guarantee with high probability. Note that this reduction holds even for an affine noise oracle $g$ where

$$\Pr(\|g(w) - \mathbb{E}[g(w)]\| \ge t) \le 2\exp(-t^2/(\sigma_0^2 + \sigma_1^2\|\mathbb{E}[g(w)]\|^2)) \tag{10}$$

by creating a new gradient oracle $\widetilde{g}(w)$ such that

$$\|\widetilde{g}(w) - \mathbb{E}[\widetilde{g}(w)]\|^2 \le 9\sigma_0^2 \ln(4T/\delta') + 9\sigma_1^2 \ln(4T/\delta')\|\mathbb{E}[\widetilde{g}(w)]\|^2.$$

More formally, under a sub-Gaussian affine noise assumption, we derive the following two results using Theorems 1 and 2:

**Theorem 3.** *Assume that $f$ is $\beta$-smooth and a sub-Gaussian gradient oracle (Equation (10)). Then for AdaSGD, for any $\delta \in (0, 1/3)$, it holds with probability at least $1 - 2\delta$ that*

*(i)* $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 \le O\Big(\frac{2F}{\eta} + \eta\beta\Big)\frac{\sigma_0\sqrt{\log\frac{T}{\delta}}}{\sqrt{T}} + O\Big(\frac{\gamma}{\eta}\Delta_1 + (1 + \sigma_1^2 \log\frac{1}{\delta}\log\frac{T}{\delta})\big(\frac{2F}{\eta} + \eta\beta\big)^2 + \sigma_0^2\log\frac{1}{\delta}\log\frac{T}{\delta}\Big)\frac{1}{T}$;

*(ii) and $f(w_t) - f^\star \le F$ for all $1 \le t \le T + 1$,*

*where $\Delta_1 = f(w_1) - f^\star$ and*

$$F = 2\Delta_1 + \big(9\log\tfrac{T}{\delta} + 12C_1\big)\eta\sigma_0\sqrt{\ln\tfrac{4T}{\delta}} + \big(81\log^2\tfrac{T}{\delta} + 144C_1^2\big)\eta^2\beta\sigma_1^2\ln\tfrac{4T}{\delta} + \eta^2\beta C_1, \tag{11}$$

$$C_1 = \log\left(1 + \frac{18\sigma_0^2 T\ln\frac{4T}{\delta} + (8 + 72\sigma_1^2\ln\frac{4T}{\delta})(\eta^2\beta^2T^3 + \beta\Delta_1 T)}{\gamma^2}\right). \tag{12}$$

**Theorem 4.** *Assume that $f$ is convex and $\beta$-smooth and a sub-Gaussian gradient oracle (Equation (10)). Then for AdaSGD, for any $\delta \in (0, 1/4)$, it holds with probability at least $1 - 2\delta$ that*

*(i)* $f\Big(\frac{1}{T}\sum_{t=1}^{T} w_t\Big) - f^\star \le O\Big(\Big(\frac{D^2}{\eta} + \eta\Big) + D\sqrt{\log\frac{1}{\delta}}\Big)\frac{\sigma_0\sqrt{\log\frac{T}{\delta}}}{\sqrt{T}}$
$+ O\Big(\frac{\gamma d_1^2}{\eta} + \beta(1 + \sigma_1^2\log\frac{T}{\delta})\Big(\frac{D^2}{\eta} + \eta\Big)^2 + \beta\sigma_1^2 D^2\log\frac{1}{\delta}\log\frac{T}{\delta}\Big)\frac{1}{T}$;

*(ii) and $\|w_t - w^\star\| \le D$ for all $1 \le t \le T + 1$,*

*where $d_1 = \|w_1 - w^\star\|$, and*

$$D^2 = \widetilde{O}\Big(d_1^2 + \eta^2\big(1 + \sigma_1^2 + \gamma^{-2}(\sigma_0^2 + \beta\sigma_1^2 F) + \gamma^{-4}(\sigma_0^4 + \beta^2\sigma_1^4F^2)\big)\Big),$$
$$F = \widetilde{O}\Big(f(w_1) - f^\star + \eta\sigma_0 + \eta^2\beta\sigma_1^2 + \eta^2\beta\Big).$$

The theorems share the same bounds with Theorems 1 and 2 up to logarithmic factors. Note that a careful modification of the analyses can yield slightly tighter bounds as Lemma 5 already holds for a sub-Gaussian martingale difference sequence.

*Proof of Theorems 3 and 4.* Let $\widetilde{g}(w)$ be the gradient oracle defined above by applying Lemma 15 to the noise of the sub-Gaussian oracle. Due to Lemma 15 this oracle returns unbiased gradients with bounded affine noise for parameters $\widetilde{\sigma}_0 = 3\sigma_0\sqrt{\ln(4T/\delta')}$ and $\widetilde{\sigma}_1 = 3\sigma_1\sqrt{\ln(4T/\delta')}$. Furthermore, with probability $1 - \delta'$, AdaSGD performs the exact steps with both oracles as the noisy gradients are equal. Hence we can use the guarantees of Theorems 1 and 2. We conclude by plugging $\widetilde{\sigma}_0$ and $\widetilde{\sigma}_1$ to the bounds and performing a union-bound with the probability the two oracles output the same noisy gradients. □

### C.1. Proof of Lemma 15

*Proof.* Denote $r = \sigma\sqrt{\ln(4/\delta)}$; note that $\Pr(\|X\| > r) \le \delta/2$ (due to Equation (9)). Define a random vector $Z$ by rejection sampling $X$ on the $r$-unit ball: if $\|X\| \le r$, let $Z = X$; otherwise, continue resampling from the distribution of $X$ and reject samples as long as their norm is larger than $r$; let $Z$ be the first accepted sample. To correct the bias of $Z$, let

$$\bar{X} = \begin{cases} Z & \text{with prob. } 1 - \delta/2; \\ -\frac{2-\delta}{\delta}\mathbb{E}[Z] & \text{with prob. } \delta/2. \end{cases}$$

By construction, $\mathbb{E}[\bar{X}] = 0$. Also, $\Pr(\bar{X} = X) \ge 1 - \delta$ due to a union bound of $\Pr(\bar{X} = Z)$ and $\Pr(X = Z)$, each happens with probability at least $1 - \delta/2$. We are left with bounding $\bar{X}$. As $\|Z\| \le r$ it remains to bound the correction term. For this,

we first bound $\mathbb{E}[Z] = \mathbb{E}[X \mid \|X\| \le r]$. By the law of total expectation,

$$\begin{aligned}\mathbb{E}[X\mathbf{1}\{\|X\| \le r\}] &= \Pr(\|X\| \le r)\mathbb{E}[X\mathbf{1}\{\|X\| \le r\} \mid \|X\| \le r] + \Pr(\|X\| > r)\mathbb{E}[X\mathbf{1}\{\|X\| \le r\} \mid \|X\| > r] \\ &= \Pr(\|X\| \le r)\mathbb{E}[X\mathbf{1}\{\|X\| \le r\} \mid \|X\| \le r] \\ &= \Pr(\|X\| \le r)\mathbb{E}[X \mid \|X\| \le r].\end{aligned}$$

Hence, as $\mathbb{E}[X] = 0$,

$$\mathbb{E}[X \mid \|X\| \le r] = \frac{\mathbb{E}[X\mathbf{1}\{\|X\| \le r\}]}{\Pr(\|X\| \le r)} = -\frac{\mathbb{E}[X\mathbf{1}\{\|X\| > r\}]}{\Pr(\|X\| \le r)}.$$

Thus, using the tail bound (Equation (9)),

$$\begin{aligned}\|\mathbb{E}[Z]\| &\le 2\mathbb{E}[\|X\|\mathbf{1}\{\|X\| > r\}] && (\Pr(\|X\| \le r) > 1/2) \\ &\le 2\int_0^\infty \Pr(\|X\|\mathbf{1}\{\|X\| > r\} \ge x)dx && \text{(tail formula for expectations)} \\ &\le 2r\Pr(\|X\| > r) + 2\int_r^\infty \Pr(\|X\|\mathbf{1}\{\|X\| > r\} \ge x)dx && \text{(split integral at } r) \\ &\le \delta r + 2\int_r^\infty \Pr(\|X\| \ge x)dx && (\Pr(\|X\| > r) \le \delta/2) \\ &\le \delta r + 4\int_r^\infty e^{-x^2/\sigma^2}dx && \text{(Equation (9))} \\ &\le \delta r + \frac{2\sigma^2}{r}\int_r^\infty \frac{2x}{\sigma^2}e^{-x^2/\sigma^2}dx && (x/r \ge 1 \text{ for } x \ge r) \\ &= \delta r - \frac{2\sigma^2}{r}e^{-x^2/\sigma^2}\Big|_r^\infty = \delta r + \frac{2\sigma^2}{r}e^{-r^2/\sigma^2} \\ &\le \delta r + \frac{\delta\sigma}{2}.\end{aligned}$$

We conclude that $\bar{X}$ is bounded in absolute value (with probability one) by

$$\max\{r, (r + \sigma/2)(2 - \delta)\} \le 3\sigma\sqrt{\ln(4/\delta)}.$$ □

## D. Lower Bound for the Effective Diameter in the Convex Case

In Section 4 we establish a bound of $\max_{t \le T+1}\|w_t - w^\star\|^2$ which has polynomial dependence in $1/\gamma$. In this section we construct a one-dimensional example which shows that such polynomial dependence is necessary in order to obtain a bound of $\max_{t \le T+1}\|w_t - w^\star\|^2$ which is $o(T)$ with high probability.

Let $f(w) = \frac{\beta}{2}w^2$ for some $0 < \beta < \frac{\sigma}{T^{3/2}}$ ($\beta$ is extremely small with respect to $\sigma$). Let $g(w)$ be a stochastic oracle such that

$$g(w) = \begin{cases}\nabla f(w) + \sigma & \text{w.p. } \frac{1}{T}; \\ \nabla f(w) - \frac{\sigma}{T-1} & \text{w.p. } 1 - \frac{1}{T},\end{cases}$$

where $T > 1$. Hence, $E[g(w) - \nabla f(w)] = 0$ and $\|g(w) - \nabla f(w)\|^2 \le \sigma^2$ with probability 1. Let $w_1 = w^\star = 0$ and $\eta = 1$. With constant probability ($\approx e^{-1}$), $g_t = \nabla f(w_t) - \sigma/(T-1)$ for all $t \in [T]$. As $\max_{t \le T+1}\|w_t - w^\star\|^2 = o(T)$ with high probability, we assume that for large enough $T$, $|w_t - w^\star| < \frac{\sqrt{T}}{2}$ for all $t \le T+1$ (still with constant probability under a union bound). Hence, with constant probability,

$$w_{T+1} = -\sum_{t=1}^T \eta_t g_t = -\sum_{t=1}^T \eta_t\Big(\beta w_t - \frac{\sigma}{T-1}\Big) \ge \frac{\sigma}{2(T-1)}\sum_{t=1}^T \eta_t,$$

where the last inequality is the assumption of small $\beta$. Thus, as $\|g_s\| = \|\beta w_t - \sigma/(T-1)\| \le \sigma/(T-1)$ (as $w_t \ge 0$ for all $t$ since the stochastic gradients are always negative),

$$\sum_{t=1}^{T} \eta_t = \sum_{t=1}^{T} \frac{1}{\sqrt{\gamma^2 + \sum_{s=1}^{t} \|g_s\|^2}} \ge \sum_{t=1}^{T} \frac{1}{\sqrt{\gamma^2 + \frac{\sigma^2 t}{(T-1)^2}}} = \frac{T-1}{\sigma} \sum_{t=1}^{T} \frac{1}{\sqrt{\frac{\gamma^2 (T-1)^2}{\sigma^2} + t}}.$$

Returning to $w_{T+1}$, bounding sum by integration,

$$w_{T+1} \ge \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sqrt{\frac{\gamma^2 (T-1)^2}{\sigma^2} + t}} \ge \sqrt{\frac{\gamma^2 (T-1)^2}{\sigma^2} + T + 1} - \sqrt{\frac{\gamma^2 (T-1)^2}{\sigma^2} + 1} = \Omega(\sigma/\gamma).$$

Thus, $\|w_{T+1} - w^\star\|^2 = \Omega(\sigma^2/\gamma^2)$.

## E. Non-convex SGD with known parameters

In this section we provide a high probability convergence analysis of Stochastic Gradient Descent (SGD) for a $\beta$-smooth objective function $f$ under an affine noise model with known parameters $(\beta, \sigma_0, \sigma_1)$. The tuned step-size we use is the following,

$$\eta = \min\left\{ \frac{1}{4\beta(1 + \sigma_1^2 \log \frac{T}{\delta})}, \frac{\alpha}{\sigma_0 \sqrt{T}} \right\}, \tag{13}$$

for some parameter $\alpha > 0$. Following is the convergence guarantee.

**Theorem 5.** *Assume that $f$ is $\beta$-smooth. Then for SGD with step-size (Equation (13)), for any $\delta \in (0, \frac{1}{2})$, it holds with probability at least $1 - 2\delta$ that*

*(i)* $\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(w_t)\|^2 \le \left(\beta\alpha + \frac{\Delta_1}{\alpha}\right) \frac{2\sigma_0}{\sqrt{T}} + \left(8\beta\Delta_1 (1 + 4\sigma_1^2) \log \frac{T}{\delta} + 24\sigma_1^2 \beta^2 \alpha^2 \log \frac{1}{\delta} + 15\sigma_0^2 \log \frac{1}{\delta}\right) \frac{1}{T}$;

*(ii) and $f(w_t) - f^\star \le F$ for all $1 \le t \le T + 1$,*

*where $\Delta_1 = f(w_1) - f^\star$ and*

$$F = 2\Delta_1 + 2\beta\alpha^2 + 3 \min\left\{ \frac{\sigma_0^2}{4\beta(1 + \sigma_1^2 \log \frac{T}{\delta})}, \frac{\sigma_0 \alpha}{\sqrt{T}} \right\} \log \frac{T}{\delta}. \tag{14}$$

Similarly to Theorem 1, Theorem 5 establishes a high probability bound of $f(w_t) - f^\star$ for all $t$. Also note that tuning of $\alpha = \Theta(\sqrt{\Delta_1/\beta})$ yields a rate of $\widetilde{O}\big(\sigma_0 \sqrt{\beta\Delta_1}/\sqrt{T} + (1 + \sigma_1^2)\beta\Delta_1/T\big)$, matching the results in expectation of Ghadimi & Lan (2013) and Bottou et al. (2018, Theorem 4.8) up to few additional logarithmic factors being a result of the high probability analysis.

### E.1. Proof of Theorem 5

In order to prove Theorem 5 the next lemma establish a bound of $f(w_t) - f^\star$ for all $t \le T$ that holds with high probability, similarly to Lemma 2 (proof given later in the section).

**Lemma 16.** *With probability at least $1 - \delta$, it holds that for all $t \le T + 1$, $f(w_t) - f^\star \le F$ (with $F$ defined in Theorem 5 ).*

We continue to prove the main result using similar lines to the proof of Theorem 1.

*Proof of Theorem 5.* We start with a regret analysis. Using $\beta$-smoothness, we have for all $t \ge 1$ that

$$f(w_{t+1}) \le f(w_t) - \eta \nabla f(w_t) \cdot g_t + \tfrac{1}{2}\beta\eta^2 \|g_t\|^2,$$

which implies

$$\nabla f(w_t) \cdot g_t \le \frac{1}{\eta}\big(f(w_t) - f(w_{t+1})\big) + \tfrac{1}{2}\beta\eta\|g_t\|^2.$$

Summing the above over $t = 1, \dots, T$, we obtain (with $\Delta_1 = f(w_1) - f^\star$)

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t \le \frac{\Delta_1}{\eta} + \frac{\beta\eta}{2}\sum_{t=1}^{T}\|g_t\|^2.$$

Observe that

$$\|g_t\|^2 \le 2\|\nabla f(w_t)\|^2 + 2\|g_t - \nabla f(w_t)\|^2 \le 2(1+\sigma_1^2)\|\nabla f(w_t)\|^2 + 2\sigma_0^2.$$

Hence,

$$\sum_{t=1}^{T}\|g_t\|^2 \le 2\sum_{t=1}^{T}(1+\sigma_1^2)\|\nabla f(w_t)\|^2 + 2\sigma_0^2 T.$$

By the definition of $\eta$ (Equation (13)),

$$\begin{aligned}\sum_{t=1}^{T} \nabla f(w_t) \cdot g_t &\le \frac{\Delta_1}{\eta} + \beta\eta\sum_{t=1}^{T}(1+\sigma_1^2)\|\nabla f(w_t)\|^2 + \beta\eta\sigma_0^2 T \\ &\le 4\beta\Delta_1(1+\sigma_1^2\log\tfrac{T}{\delta}) + \frac{1}{4}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 + \sigma_0\sqrt{T}\Big(\beta\alpha + \frac{\Delta_1}{\alpha}\Big). \end{aligned} \tag{15}$$

We continue with a martingale analysis to replace $\sum_{t=1}^{T}\nabla f(w_t)\cdot g_t$ with $\sum_{t=1}^{T}\|\nabla f(w_t)\|^2$. Consider the sequence of R.V.s $(Z_t)_{t=1}^{T}$ defined by $Z_t = \nabla f(w_t)\cdot(\nabla f(w_t) - g_t)$. Then $(Z_t)_{t=1}^{T}$ is a martingale difference sequence, as

$$\mathbb{E}_{t-1}[Z_t] = \nabla f(w_t)\cdot(\nabla f(w_t) - \mathbb{E}_{t-1}[g_t]) = 0.$$

By Cauchy-Schwarz inequality and the noise assumption,

$$|Z_t| \le \|\nabla f(w_t)\|\|\nabla f(w_t) - g_t\| \le \|\nabla f(w_t)\|\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2}.$$

Thus, we obtain from Lemma 5 with

$$|Z_t| \le \|\nabla f(w_t)\|\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2}$$

and $\lambda = 1\big/3(\sigma_0^2 + 2\beta\sigma_1^2 F)$ that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\nabla f(w_t)\cdot(\nabla f(w_t) - g_t) \le \frac{\sum_{t=1}^{T}(\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2)\|\nabla f(w_t)\|^2}{4(\sigma_0^2 + 2\beta\sigma_1^2 F)} + 3(\sigma_0^2 + 2\beta\sigma_1^2 F)\log(1/\delta).$$

From Lemmas 8 and 16, with probability at least $1-\delta$, $\|\nabla f(w_t)\|^2 \le 2\beta F$. Hence, under a union bound with Lemma 16, with probability at least $1-2\delta$,

$$\sum_{t=1}^{T}\nabla f(w_t)\cdot(\nabla f(w_t) - g_t) \le \frac{1}{4}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 + 3(\sigma_0^2 + 2\beta\sigma_1^2 F)\log\tfrac{1}{\delta}. \tag{16}$$

Summing Equations (15) and (16) and rearranging,

$$\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 \le 8\beta\Delta_1(1+\sigma_1^2\log\tfrac{T}{\delta}) + 6(\sigma_0^2 + 2\beta\sigma_1^2 F)\log\tfrac{1}{\delta} + 2\sigma_0\sqrt{T}\Big(\beta\alpha + \frac{\Delta_1}{\alpha}\Big).$$

From Lemma 16 and Equation (13),

$$F \le 2\Delta_1 + 2\beta\alpha^2 + \frac{3\sigma_0^2}{4\beta\sigma_1^2}.$$

Thus,

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 &\le 8\beta\Delta_1(1+\sigma_1^2\log\tfrac{T}{\delta}) + 24\beta\sigma_1^2(\Delta_1+\beta\alpha^2)\log\tfrac{1}{\delta} + 15\sigma_0^2\log\tfrac{1}{\delta} + 2\sigma_0\sqrt{T}\Big(\beta\alpha + \frac{\Delta_1}{\alpha}\Big)\\ &\le 8\beta\Delta_1(1+4\sigma_1^2)\log\tfrac{T}{\delta} + 24\sigma_1^2\beta^2\alpha^2\log\tfrac{1}{\delta} + 15\sigma_0^2\log\tfrac{1}{\delta} + 2\sigma_0\sqrt{T}\Big(\beta\alpha + \frac{\Delta_1}{\alpha}\Big).\end{aligned}$$

□

### E.2. Proof of Lemma 16

*Proof.* Consider the sequence of RVs $(Z_t)_{t=1}^T$ defined by $Z_t = \nabla f(w_t)\cdot(\nabla f(w_t) - g_t)$. Then $(Z_t)_{t=1}^T$ is a martingale difference sequence, as

$$\mathbb{E}_{t-1}[Z_t] = \nabla f(w_t)\cdot(\nabla f(w_t) - \mathbb{E}_{t-1}[g_t]) = 0.$$

From Cauchy–Schwarz inequality and the noise model, $|Z_t| \le \|\nabla f(w_t)\|\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2}$. Thus, we obtain from Lemma 5 (together with a union bound over $t$) that for any fixed $\lambda > 0$, with probability at least $1-\delta$, for all $1 \le t \le T$,

$$\sum_{s=1}^{t}\nabla f(w_s)\cdot(\nabla f(w_s) - g_s) \le \frac{3\lambda}{4}\sum_{s=1}^{t}\|\nabla f(w_s)\|^2(\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2) + \frac{\log\frac{T}{\delta}}{\lambda}.$$

We will proceed by induction. The claim is immediate for $t = 1$ and we move to prove it for $t+1$. Using smoothness,

$$f(w_{s+1}) \le f(w_s) - \eta\nabla f(w_s)\cdot g_s + \frac{\beta\eta^2}{2}\|g_s\|^2.$$

Summing the above over $s = 1,\dots,t$,

$$f(w_{t+1}) - f(w_1) \le -\eta\sum_{s=1}^{t}\nabla f(w_s)\cdot g_s + \frac{\beta\eta^2}{2}\sum_{s=1}^{t}\|g_s\|^2. \tag{17}$$

The second term of the right hand side can be bounded by the definitions of $\eta$ and the noise model,

$$\begin{aligned}\frac{\beta\eta^2}{2}\sum_{s=1}^{t}\|g_s\|^2 &\le \beta\eta^2\sum_{s=1}^{t}\big(\|g_s - \nabla f(w_s)\|^2 + \|\nabla f(w_s)\|^2\big) && (\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2)\\ &\le \beta\eta^2\sum_{s=1}^{t}\big((1+\sigma_1^2)\|\nabla f(w_s)\|^2 + \sigma_0^2\big) && (\|\nabla f(w_s) - g_s\|^2 \le \sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2)\\ &\le \beta\alpha^2 + \frac{\eta}{4}\sum_{s=1}^{t}\|\nabla f(w_s)\|^2. && \text{(Equation (13))}\end{aligned}$$

We move to bound the first term of the right hand side. We split the sum to obtain a martingale term,

$$-\eta\sum_{s=1}^{t}\nabla f(w_s)\cdot g_s = -\eta\sum_{s=1}^{t}\|\nabla f(w_s)\|^2 + \eta\sum_{s=1}^{t}\nabla f(w_s)\cdot(\nabla f(w_s) - g_s).$$

From Lemma 8 and our induction, $\|\nabla f(w_t)\|^2 \le 2\beta(f(w_t) - f^\star) \le 2\beta F$. Returning to our martingale bound, we have

$$\begin{aligned}\sum_{s=1}^{t}\nabla f(w_s)\cdot(\nabla f(w_s) - g_s) &\le \frac{3\lambda}{4}\sum_{s=1}^{t}\|\nabla f(w_s)\|^2(\sigma_0^2 + \sigma_1^2\|\nabla f(w_s)\|^2) + \frac{\log\frac{T}{\delta}}{\lambda}\\ &\le \frac{3\lambda}{4}(\sigma_0^2 + 2\beta\sigma_1^2 F)\sum_{s=1}^{t}\|\nabla f(w_s)\|^2 + \frac{\log\frac{T}{\delta}}{\lambda}.\end{aligned}$$

Setting $\lambda = 1 \Big/ (\sigma_0^2 + 2\beta\sigma_1^2 F)$,

$$\sum_{s=1}^{t} \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) \leq \frac{3}{4} \sum_{s=1}^{t} \|\nabla f(w_s)\|^2 + (\sigma_0^2 + 2\beta\sigma_1^2 F) \log \tfrac{T}{\delta}.$$

Returning to Equation (17) and combining the inequalities,

$$\begin{aligned} f(w_{t+1}) - f(w_1) &\leq \beta\alpha^2 + \eta(\sigma_0^2 + 2\beta\sigma_1^2 F) \log \tfrac{T}{\delta} \\ &\leq \beta\alpha^2 + \eta\sigma_0^2 \log \tfrac{T}{\delta} + \frac{1}{2} F = F - (f(w_1) - f^\star), \end{aligned}$$

where the last equality follows from the definition of $\eta$. Adding $f(w_1) - f^\star$ to both sides concludes the proof. □