Proceedings of Machine Learning Research vol 195:1–25, 2023 36th Annual Conference on Learning Theory

Breaking the Lower Bound with (Little) Structure: Acceleration in Non-Convex Stochastic Optimization with Heavy-Tailed Noise

Zijian Liu ZL3067@STERN.NYU.EDU
Jiawei Zhang JZHANG@STERN.NYU.EDU
Zhengyuan Zhou ZZHOU@STERN.NYU.EDU
Stern School of Business, New York University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

In this paper, we consider the stochastic optimization problem with smooth but not necessarily convex objectives in the heavy-tailed noise regime, where the stochastic gradient's noise is assumed to have bounded pth moment ($p \in (1,2]$). This is motivated by a recent plethora of studies in the machine learning literature, which point out that, in comparison to the standard finite-variance assumption, the heavy-tailed noise regime is more appropriate for modern machine learning tasks such as training neural networks. In the heavy-tailed noise regime, Zhang et al. (2020) is the first to prove the $\Omega(T^{\frac{1-p}{3p-2}})$ lower bound for convergence (in expectation) and provides a simple clipping algorithm that matches this optimal rate. Later, Cutkosky and Mehta (2021) proposes another algorithm, which is shown to achieve the nearly optimal high-probability convergence guarantee $O(\log(T/\delta)T^{\frac{1-p}{3p-2}})$, where δ is the probability of failure. However, this desirable guarantee is only established under the additional assumption that the stochastic gradient itself is bounded in pth moment, which fails to hold even for quadratic objectives and centered Gaussian noise.

In this work, we first improve the analysis of the algorithm in Cutkosky and Mehta (2021) to obtain the same nearly optimal high-probability convergence rate $O(\log(T/\delta)T^{\frac{1-p}{3p-2}})$, without the above-mentioned restrictive assumption. Next, and curiously, we show that one can achieve a faster rate than that dictated by the lower bound $\Omega(T^{\frac{1-p}{3p-2}})$ with only a tiny bit of structure, i.e., when the objective function $F(x)$ is assumed to be in the form of $\mathbb{E}_{\Xi\sim\mathcal{D}}[f(x,\Xi)]$, arguably the most widely applicable class of stochastic optimization problems. For this class of problems, we propose the first variance-reduced accelerated algorithm and establish that it guarantees a high-probability convergence rate of $O(\log(T/\delta)T^{\frac{1-p}{2p-1}})$ under a mild condition, which is faster than $\Omega(T^{\frac{1-p}{3p-2}})$. Notably, even when specialized to the standard finite-variance case ($p=2$), our result yields the (near-)optimal high-probability rate $O(\log(T/\delta)T^{-1/3})$, which is unknown before.

Keywords: Non-convex stochastic optimization, variance reduction, accelerated algorithm

1. Introduction

In this paper, we consider the optimization problem with the objective function $F(x):\mathbb{R}^d\to\mathbb{R}$, where $F(x)$ is smooth but not necessarily convex. With a gradient oracle (i.e. $\nabla F(x)$ accessible at every $x\in\mathbb{R}^d$), it is well-known that Gradient Descent (GD) algorithm converges in the rate of $\Theta(T^{-1/2})$ after T iterations for finding the critical point. However, in practical scenarios, (at best) an unbiased noisy estimate $\widehat{\nabla}F(x)$ ($\mathbb{E}[\widehat{\nabla}F(x)|x]=\nabla F(x)$) is available. In such cases, the gold standard is Stochastic Gradient Descent (SGD), a classical first-order method that has been widely deployed for modern machine learning tasks (such as training deep neural networks). Motivated by its empirical success, Ghadimi and Lan (2013) has characterized its theoretical guarantees for non-convex objectives and is the first to establish that SGD converges in expectation with the rate of $O(T^{-1/4})$ for minimizing the gradient norm under the standard finite-variance assumption. The rate $O(T^{-1/4})$ is also shown to be optimal without any additional assumptions (Arjevani et al., 2019).

However, recent studies (Simsekli et al., 2019; Şimşekli et al., 2019; Zhang et al., 2020) point out that assuming finite-variance noise is too optimistic for modern machine learning tasks (in particular, for training neural networks) and it is more appropriate to assume that the noise only has bounded pth moment, i.e., $\mathbb{E}[\|\widehat{\nabla} F(x)-\nabla F(x)\|^p|x] \leq \sigma^p$ for some $p \in (1,2]$, which is known as the heavy-tailed regime. This brings significant challenges in algorithmic and theoretical fronts, as SGD may fail to converge and the existing theory for SGD becomes invalid when $p \neq 2$.

Recently, exciting new developments have made progress in overcoming these challenges. In particular, Zhang et al. (2020); Cutkosky and Mehta (2021) propose two new provable algorithms, both of which employ the clipping gradient method. More precisely, given a stochastic gradient $\widehat{\nabla} F(x)$, they both consider a new truncated random variable $g = \min\left\{1, \frac{M}{\|\widehat{\nabla} F(x)\|}\right\} \widehat{\nabla} F(x)$ where M is the clipping magnitude, in replacement of using $\widehat{\nabla} F(x)$ directly. As long as M is picked appropriately, Zhang et al. (2020) shows a convergence rate (in expectation) of $O(T^{\frac{1-p}{3p-2}})$ for SGD combined with clipping directly, which matches the (in-expectation) lower bound $\Omega(T^{\frac{1-p}{3p-2}})$ proved in the same paper. A step further, Cutkosky and Mehta (2021) studies the high-probability convergence behavior, which provides a stronger guarantee for each individual run. Their (modified) SGD with clipping is shown to converge at a rate of $O(\log(T/\delta)T^{\frac{1-p}{3p-2}})$ with probability at least $1-\delta$. Unfortunately, the result in Cutkosky and Mehta (2021) requires that the stochastic gradients are bounded in pth moment: $\mathbb{E}[\|\widehat{\nabla} F(x)\|^p|x] \leq G^p$ for some $G > 0$, which is rather restrictive since it can not even hold when $F(x)$ is considered as quadratic and $\widehat{\nabla} F(x) - \nabla F(x)$ is an independent centered Gaussian random variable (note that the tail of the Gaussian noise is very light). As such, when only considering the standard bounded pth moment noise assumption, we are naturally led to the following question:

Q1: Is it possible to design an algorithm with a provable high-probability convergence guarantee that (nearly) matches the lower bound $\Omega(T^{\frac{1-p}{3p-2}})$ for the general class of problems?

Moving beyond the general stochastic optimization problem, a particular – but still general enough – subclass of problems that are of special interest are $F(x) = \mathbb{E}_{\Xi\sim\mathcal{D}}[f(x,\Xi)]$, where $f(x,\Xi)$ is assumed to be differentiable with respect to x for every realization of Ξ drawn from a (possibly unknown) probability distribution $\mathcal{D}$. In general, $\nabla f(x,\Xi)$ is an unbiased estimator of $\nabla F(x)$: $\mathbb{E}_{\Xi\sim\mathcal{D}}[\nabla f(x,\Xi)|x] = \nabla F(x)$. This structure has attracted significant attention from the optimization community as many modern machine learning problems can be formulated in such a form. A recent breakthrough to improve the performance of algorithms for solving this class of problems is to add the variance reduction, which is shown to achieve acceleration. More specifically, under the finite-variance ($p = 2$) and additional averaged smoothness assumptions (i.e. $\mathbb{E}_{\Xi\sim\mathcal{D}}[\|\nabla f(x,\Xi) - \nabla f(y,\Xi)\|^2] \leq L^2\|x-y\|^2, \forall x,y \in \mathbb{R}^d$), Fang et al. (2018); Cutkosky and Orabona (2019); Tran-Dinh et al. (2019); Liu et al. (2020); Li et al. (2021) propose different algorithms with the same convergence rate of $O(T^{-1/3})$ in expectation, which matches the lower bound in Arjevani et al. (2019) and is also faster than the generic $\Theta(T^{-1/4})$ for SGD.

However, it still remains open whether – and if so, how – similar acceleration can be achieved in the presence of heavy-tailed noise when $p \in (1,2)$. If the convergence rate as a function of p is continuous (a big “if" that by no means holds *a priori*), then one would have some hope to do better than the general lower bound $\Omega(T^{\frac{1-p}{3p-2}})$, since when $p = 2$, $\Omega(T^{\frac{1-p}{3p-2}})$ yields $\Omega(T^{-1/4})$, which we know can be improved to $\Theta(T^{-1/3})$ in the subclass. Whereas, it is unclear which acceleration scheme – if any – would be effective in the heavy-tailed noise setting (under the subclass of the problems of the form $F(x) = \mathbb{E}_{\Xi\sim\mathcal{D}}[f(x,\Xi)]$), thereby leading to the second main question:

Q2: Is it possible to find an algorithm with a provable convergence guarantee that outperforms the general lower bound $\Omega(T^{\frac{1-p}{3p-2}})$ and (nearly) matches the optimal $\Theta(T^{-1/3})$ rate when $p = 2$?

1.1. Our Contributions

We provide affirmative answers to both questions. For *Q1*, surprisingly, the algorithm in Cutkosky and Mehta (2021) without any modification is enough: under an improved analysis, the additional assumption of bounded pth moment stochastic gradients can be removed. To do so, we revisit the algorithm Normalized SGD with Clipping and Momentum, proposed in Cutkosky and Mehta (2021) and improve the analysis by employing the proof idea from Gorbunov et al. (2020) to obtain a better result. To be more precise, *without* the assumption of bounded pth moment stochastic gradients and therefore for the general class of non-convex stochastic optimization problems, we show that the algorithm can converge at the rate of $O(\log(T/\delta)T^{\frac{1-p}{3p-2}})$ with probability at least $1-\delta$.

For *Q2*, we provide a new and *the first accelerated* algorithm for the heavy-tailed noise setting and establish the convergence rate of $O(\log(T/\delta)T^{\frac{1-p}{2p-1}})$ with probability at least $1-\delta$ under a mild condition, which is faster than the general lower bound $\Omega(T^{\frac{1-p}{3p-2}})$ and reduces to the nearly optimal rate $O(\log(T/\delta)T^{-1/3})$ when $p = 2$. Our algorithm is designed by integrating a new variant of the variance-reduced gradient estimator (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019; Liu et al., 2020) into the Normalized SGD with Clipping and Momentum algorithm. To the best of our knowledge, this is the first algorithm provably guaranteeing a faster convergence rate compared with the existing lower bound $\Omega(T^{\frac{1-p}{3p-2}})$ that is proved for the general heavy-tailed noise problem. When specialized to $p = 2$, our result yields the nearly optimal high-probability $\widetilde{\Theta}(T^{-1/3})$ rate for the standard finite-variance setting, thereby improving the existing state of knowledge where only in-expectation bound (of the same rate) is known.

1.2. Related Work

Convergence with heavy-tailed noise: When the noise is assumed to have the finite pth moment for $p \in (1,2]$, Zhang et al. (2020) is the first to prove an $O(T^{\frac{1-p}{3p-2}})$ convergence rate in expectation by combining SGD and clipping directly. Later, Cutkosky and Mehta (2021) proposes a novel algorithm, Normalized SGD with Clipping and Momentum, which enjoys the provable high-probability convergence behavior attaining the rate of $O(\log(T/\delta)T^{\frac{1-p}{3p-2}})$ with probability at least $1-\delta$ after T iterations running. Compared with Zhang et al. (2020), the convergence rate is almost the same up to an extra logarithmic factor, whereas, Cutkosky and Mehta (2021) requires the additional restrictive assumption of bounded pth moment gradient estimators. Gorbunov et al. (2020) is the first to prove the high-probability bounds of clipping algorithms for smooth convex optimization on $\mathbb{R}^d$ when $p = 2$. We extend their proof idea in this work.

Lower bound with heavy-tailed noise: As far as we know, the only existing lower bound when considering the heavy-tailed noise appears in Zhang et al. (2020), in which the authors prove that the convergence rate of any algorithm for finding the critical point can not exceed $\Omega(T^{\frac{1-p}{3p-2}})$ when the objective function is assumed to be smooth. This means that the results in Zhang et al. (2020); Cutkosky and Mehta (2021) are both (nearly) optimal. However, if $F(x)$ admits the special structure $F(x) = \mathbb{E}_{\Xi\sim\mathcal{D}}[f(x,\Xi)]$ and satisfies the averaged smoothness property, the lower bound $\Omega(T^{-1/4})$ is not tight anymore when $p = 2$. Arjevani et al. (2019) provides an improved result $\Omega(T^{-1/3})$ for this special case. But if p is considered as strictly smaller than 2, whether a tighter lower bound exists or not remains unknown.

Variance reduction for stochastic optimization: Roux et al. (2012); Johnson and Zhang (2013); Shalev-Shwartz and Zhang (2013); Mairal (2013); Defazio et al. (2014) first introduce the variance reduction technique to speed up the convergence when the objective function is convex and defined in the finite-sum form. After the important intermediate work of Allen-Zhu (2017), Lan et al. (2019); Zhou et al. (2019); Song et al. (2020); Liu et al. (2022); Carmon et al. (2022) propose different algorithms provably attaining the near-optimal or optimal convergence rate under different situations. For non-convex problems, the variance reduction technique also has been proved to improve the convergence rate in different settings. When $F(x) = \mathbb{E}_{\Xi\sim\mathcal{D}}[f(x,\Xi)]$, a large number of works (Fang et al., 2018; Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019; Liu et al., 2020; Li et al., 2021) prove the $O(T^{-1/3})$ convergence rate in expectation for new algorithms, which improves upon the well-known speed of $\Theta(T^{-1/4})$ for the vanilla SGD or momentum SGD and matches the lower bound of $\Omega(T^{-1/3})$ (Arjevani et al., 2019) under the averaged smoothness assumption. However, to the best of our knowledge, no results have been established for heavy-tailed noises among all existing works related to variance reduction.

2. Preliminaries

Notations: Let $[d]$ denote the set $\{1, 2, \cdots, d\}$ for any integer $d \geq 1$. $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product on $\mathbb{R}^d$. $\|\cdot\|$ represents ℓ_2 norm. $a \wedge b$ and $a \vee b$ are defined as $\min\{a, b\}$ and $\max\{a, b\}$ respectively. $\mathrm{sgn}(x)$ indicates the sign function satisfying $\mathrm{sgn}(x) = 1$ for $x \geq 0$ and -1 otherwise.

We focus on the following two non-convex optimization problems in this work,

P1: We consider the following problem

$$\min_{x\in\mathbb{R}^d} F(x) \tag{1}$$

where the function $F(x)$ is only assumed to be differentiable on $\mathbb{R}^d$ but without any special structure.

P2: In this problem, our objective function is chosen to have the following special form

$$\min_{x\in\mathbb{R}^d} F(x) = \mathbb{E}_{\Xi\sim\mathcal{D}}\left[f(x,\Xi)\right] \tag{2}$$

where Ξ obeys a (possibly unknown) probability distribution $\mathcal{D}$. We will omit the writing of the distribution $\mathcal{D}$ for simplicity in the remaining paper. In this case, we assume that both $F(x)$ and $f(x,\Xi)$ are differentiable with respect to any x on $\mathbb{R}^d$. $\nabla f(x,\Xi)$ denotes the gradient taken on x for any realization Ξ drawn from the distribution $\mathcal{D}$.

We note that P1 and P2 can cover most non-convex stochastic optimization problems, hence which are very general. Additionally, our analysis relies on the following assumptions.

(1) Finite lower bound: $F_* = \inf_{x\in\mathbb{R}^d} F(x) > -\infty$.

(2) Unbiased gradient estimator: We are able to access a history-independent, unbiased gradient estimator for both P1 and P2. More specifically, for P1, a stochastic gradient estimator $\widehat{\nabla} F$ satisfying

$\mathbb{E}[\widehat{\nabla} F(x)|x] = \nabla F(x), \forall x \in \mathbb{R}^d$ is provided; for P2, we are able to draw independent Ξ from the distribution $\mathcal{D}$ and compute $\nabla f(x, \Xi)$ satisfying $\mathbb{E}[\nabla f(x, \Xi)|x] = \nabla F(x)$.

(3) Bounded pth moment noise: There exist $p \in (1, 2]$ and $\sigma \geq 0$ denoting the noise level such that $\mathbb{E}[\|\widehat{\nabla} F(x) - \nabla F(x)\|^p|x] \leq \sigma^p$ for P1 and $\mathbb{E}[\|\nabla f(x, \Xi) - \nabla F(x)\|^p|x] \leq \sigma^p$ for P2.

(4) L-smoothness: $\exists L > 0$ such that $\forall x, y \in \mathbb{R}^d$, $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$ for P1 and $\|\nabla f(x, \Xi) - \nabla f(y, \Xi)\| \leq L\|x - y\|$ with probability 1 for P2.

Here we briefly discuss our assumptions. First, Assumptions (1) and (2) are common in the related literature on stochastic optimization problems. Assumption (3) is the definition of the heavy-tailed noise, which includes the widely used finite variance assumption as a subcase by considering $p = 2$. Assumption (4) for P1 is standard for smooth optimization problems. Though the almost surely smoothness property for P2 seems stronger, it is realistic in practice and used in lots of works, e.g., Cutkosky and Orabona (2019); Levy et al. (2021). We remark that Assumption (4) for P2 implies that $F(x)$ itself is also L-smooth. In section 5, we discuss the limitation of this assumption for P2. The following two facts are well-known results under our assumptions, the proof of which can be found in Nesterov et al. (2018); Lan (2020), hence, is omitted here.

Fact 1 *Under Assumption (4), we have* $F(x) \leq F(y) + \langle \nabla F(y), x - y\rangle + \frac{L}{2}\|x - y\|^2, \forall x, y \in \mathbb{R}^d$.

Fact 2 *Under Assumptions (1) and (4), we have* $\|\nabla F(x)\| \leq \sqrt{2L(F(x) - F_*)}, \forall x \in \mathbb{R}^d$.

3. Algorithms and Results

In this section, we first state the improved result for Algorithm 1 proposed by Cutkosky and Mehta (2021) in Section 3.1. Then we present our new Algorithm 2 along with its convergence theorem in Section 3.2. Our algorithm is the first to achieve the accelerated convergence rate $O(\log(T/\delta)T^{\frac{1-p}{2p-1}})$ beyond the known lower bound $\Omega(T^{\frac{1-p}{3p-2}})$.

3.1. Improved Result for the Existing Algorithm

Algorithm 1 Normalized SGD with Clipping and Momentum (Cutkosky and Mehta, 2021)

Input: $x_1 \in \mathbb{R}^d$, $0 \leq \beta < 1$, $M > 0$, $\eta > 0$.
Set $d_0 = \mathbf{0}$
for $t = 1$ **to** T **do**
  $g_t = \left(1 \wedge \frac{M}{\|\widehat{\nabla} F(x_t)\|}\right) \widehat{\nabla} F(x_t)$
  $d_t = \beta d_{t-1} + (1 - \beta)\, g_t$
  $x_{t+1} = x_t - \eta \frac{d_t}{\|d_t\|}$
end for

We introduce every part of Algorithm 1 here briefly and refer the reader to Cutkosky and Mehta (2021) for details. Algorithm 1 integrates three main techniques: gradient clipping, momentum update and normalization. The clipped gradient is to deal with the heavy-tailed noise issue. Injecting momentum into the (clipped) stochastic gradient vector can be viewed as to correct update direction. Normalizing in the update rule allows to significantly simplify the analysis.

Our improved theoretical result is shown in Theorem 3. Unlike Cutkosky and Mehta (2021), we no more require the restrictive assumption of bounded pth moment stochastic gradients but obtain the same convergence rate, which is known to be optimal up to a logarithmic factor.

Theorem 3 *Considering P1 with Assumptions (1), (2), (3) and (4), let $\Delta_1 = F(x_1) - F_*$, then under the following choices after T iterations running*

$$\beta = 1 - T^{-\frac{p}{3p-2}}; \qquad M = \frac{\sigma}{(1-\beta)^{1/p}} \vee 4\sqrt{L\Delta_1};$$
$$\eta = \sqrt{\frac{(1-\beta)\,\Delta_1}{6TL}} \wedge \frac{1-\beta}{9\beta}\sqrt{\frac{\Delta_1}{L}} \wedge \frac{\Delta_1}{120TM\,(1-\beta)\log\frac{4T}{\delta}}.$$

Algorithm 1 guarantees that with probability at least $1-\delta$, there is

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla_t\| = O\left(\frac{\sqrt{L\Delta_1}}{T^{\frac{p-1}{3p-2}}} \vee \frac{\sigma\log\frac{T}{\delta}}{T^{\frac{p-1}{3p-2}}} \vee \frac{\sqrt{L\Delta_1}\log\frac{T}{\delta}}{T^{\frac{p}{3p-2}}}\right) = O\left(\frac{\log\frac{T}{\delta}}{T^{\frac{p-1}{3p-2}}}\right).$$

We first discuss the choices of parameters. The momentum parameter β is chosen essentially the same as in Cutkosky and Mehta (2021). However, the clipping magnitude M is very different because there no more exists a uniform upper bound G on $\mathbb{E}[\|\widehat{\nabla}F(x)\|^p]^{1/p}$ that is used to decide $M = G/(1-\beta)^{1/p}$ in Cutkosky and Mehta (2021). Intuitively, one can recognize σ as a proxy of G in the current choice of M. The appearance of the term $4\sqrt{L\Delta_1}$ in M is due to the proof technique. The interested reader could refer to Section 4 for a detailed explanation. Finally, the step size η is chosen by balancing every other term to get the right convergence rate.

We would like to talk about the convergence guarantee further. At first glance, the rate seems perfect since it already matches the lower bound $\Omega(T^{-\frac{1-p}{3p-2}})$ up to a logarithmic factor. However, the main drawback of this result is lack of adaptivity to the noise parameter σ. In other words, when $\sigma = 0$, the best rate we can obtain is still $\widetilde{O}(T^{-1/2} + T^{-1/4})$ by taking $p = 2$ (note that $p \in (1, 2]$ can be chosen arbitrarily when $\sigma = 0$), which is far from the optimal rate $\Theta(T^{-1/2})$[1]. For now, how to get a (nearly) optimal rate at the same time adapting to the level of noise σ is still unclear to us.

3.2. The First Accelerated Algorithm with Heavy-Tailed Noise

Algorithm 2 Accelerated Normalized SGD with Clipping and Momentum

Input: $x_1 \in \mathbb{R}^d$, $0 \le \beta < 1$, $M > 0$, $\eta > 0$.
Set $d_0 = \mathbf{0}$
for $t = 1$ **to** T **do**
  sample $\Xi_t \sim \mathcal{D}$
  $g_t = \left(1 \wedge \frac{M}{\|\nabla f(x_t,\Xi_t)\|}\right)\nabla f(x_t, \Xi_t)$
  $d_t = \beta d_{t-1} + (1-\beta)\,g_t + \mathbb{1}_{t\ge 2}\beta\left(\nabla f(x_t,\Xi_t) - \nabla f(x_{t-1},\Xi_t)\right)$
  $x_{t+1} = x_t - \eta\frac{d_t}{\|d_t\|}$
end for

1. We note that the bounds in Cutkosky and Mehta (2021) suffer the same issue.

We are now ready to state our new algorithm designed for P2, Accelerated Normalized SGD with Clipping and Momentum, as shown in Algorithm 2, the construction of which is mainly inspired by the works (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019; Liu et al., 2020; Cutkosky and Mehta, 2021). Compared with Algorithm 1, the key difference is in how to update the vector d_t. In the Normalized SGD with Clipping and Momentum algorithm, d_t is defined as

$$d_t = \beta d_{t-1} + (1-\beta)\, g_t, \tag{3}$$

which adds momentum to the update simply. Though it is believed that the momentum part in (3) can reduce the bias between d_t and the true gradient $\nabla F(x_t)$ to accelerate the convergence, as far as we know, no theoretical justification has been established for this guess in non-convex optimization problems even considering the case $g_t = \nabla f(x_t, \Xi_t)$.

In contrast, our gradient estimator d_t comes from the framework of momentum-based variance-reduced SGD put forward by Cutkosky and Orabona (2019); Tran-Dinh et al. (2019); Liu et al. (2020). The original template is proposed under the finite variance noise assumption and is written as follows (consider $t \geq 2$ for simplicity)

$$d_t = \beta\left(d_{t-1} - \nabla f(x_{t-1}, \Xi_t)\right) + \nabla f(x_t, \Xi_t). \tag{4}$$

However, this definition can not be applied to the heavy-tailed noise case directly. Thanks to the analysis for (4) in previous works, we know that part (ii) in (5) (reformulation of (4)) actually plays a critical role in the effect of the variance reduction. Part (i) can be thought of as the same as the traditional momentum update rule.

$$d_t = \underbrace{\beta d_{t-1} + (1-\beta)\nabla f(x_t, \Xi_t)}_{(i)} + \underbrace{\beta\left(\nabla f(x_t, \Xi_t) - \nabla f(x_{t-1}, \Xi_t)\right)}_{(ii)}. \tag{5}$$

Hence the idea for the new definition of $d_t = \beta d_{t-1} + (1-\beta)\, g_t + \beta\left(\nabla f(x_t, \Xi_t) - \nabla f(x_{t-1}, \Xi_t)\right)$ in Algorithm 2 is natural and clear now, which incorporates the momentum rule of (3) used in Algorithm 1 and part (ii) in (5) to utilize the variance reduction technique.

Next, we turn to the convergence guarantee of Algorithm 2 shown in Theorem 4. One can see the variance reduction idea indeed works and improves the convergence rate to $O(\log(T/\delta)T^{\frac{1-p}{2p-1}})$, which is faster than $\Omega(T^{\frac{1-p}{3p-2}})$ when $p \in (1, 2]$. Therefore, our algorithm is the first to achieve the acceleration in the heavy-tailed noise regime but beyond the general lower bound. Notably, when $p = 2$, the speed reduces to $O(\log(T/\delta)T^{-1/3})$ nearly matching the lower bound $\Omega(T^{-1/3})$ for P2.

Theorem 4 *Considering P2 with Assumptions (1), (2), (3) and (4), let $\Delta_1 = F(x_1) - F_*$, then under the following choices after T iterations running*

$$\beta = 1 - T^{-\frac{p}{2p-1}}; \qquad M = \frac{\sigma}{(1-\beta)^{1/p}} \vee 4\sqrt{L\Delta_1};$$

$$\eta = \sqrt{\frac{\sqrt{1-\beta}\Delta_1}{60TL\log\frac{4T}{\delta}}} \wedge \frac{1-\beta}{9\beta}\sqrt{\frac{\Delta_1}{L}} \wedge \frac{\Delta_1}{120TM(1-\beta)\log\frac{4T}{\delta}}.$$

Algorithm 2 guarantees that with probability at least $1-2\delta$, there is

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla_t\| = O\left(\frac{\sqrt{L\Delta_1\log\frac{T}{\delta}}}{T^{\frac{3p-2}{4(2p-1)}}} \vee \frac{\sqrt{L\Delta_1}}{T^{\frac{p-1}{2p-1}}} \vee \frac{\sigma\log\frac{T}{\delta}}{T^{\frac{p-1}{2p-1}}} \vee \frac{\sqrt{L\Delta_1}\log\frac{T}{\delta}}{T^{\frac{p}{2p-1}}}\right) = O\left(\frac{\log\frac{T}{\delta}}{T^{\frac{p-1}{2p-1}}}\right).$$

Finally, let us discuss Theorem 4 a bit. First, the probability $1-2\delta$ is only chosen to simplify the proof, which can be replaced by $1-\delta$ via changing every δ to $\delta/2$ in the parameters. Second, M still keeps the same as it Theorem 3. In contrast, the choices of β and η are different and more important to accelerate the convergence. In particular, the order of T in β should be chosen carefully. Besides, whether the rate is tight or not is unknown for $p \in (1,2)$. Lastly, we need to mention that Theorem 4 admits the same flaw of losing adaptivity to the noise σ as Theorem 3.

4. Theoretical Analysis

We present the ideas for proving Theorems 3 and 4 here and state some important lemmas, the omitted proofs of which are provided in Section B. The proof of Theorem 4 is delivered in the last part of this section. We defer the proof of Theorem 3 to Section C in the appendix.

Our technique contributions can be summarized as follows. We first extend the ideas used in Gorbunov et al. (2020) as mentioned, which allows us to forgo the extra assumption in Cutkosky and Mehta (2021). Due to this, both theories only rely on the standard heavy-tailed assumption. Besides, we modify the proof framework in Cutkosky and Mehta (2021) to make it compatible with Gorbunov et al. (2020) and give an almost unified analysis for Algorithms 1 and 2 together. Additionally, the effect of variance reduction in d_t in Algorithm 2 is quantified in a high-probability manner, in contrast, which is generally measured by an in-expectation bound previously. Lastly, we would like to emphasize that our proof technique can be easily extended to the same $(2, C \geq 1)$-smooth Banach space used in Cutkosky and Mehta (2021). However, we stick to $\mathbb{R}^d$ in this paper for simplicity.

Before diving into the proof, we first outline the key thoughts used in the analysis. The goal is to show that the event

$$E_\tau = \left\{ \eta \sum_{s=1}^{t} \|\nabla F(x_s)\| + F(x_{t+1}) - F_* \leq 2(F(x_1) - F_*), \forall t \leq \tau \right\} \tag{6}$$

holds with high probability for every time $0 \leq \tau \leq T$. We remark that Gorbunov et al. (2020) aims to prove the event (simplified) $\left\{\eta \sum_{s=1}^{t} F(x_s) - F(x_*) + \|x_{t+1} - x_*\|^2 \leq 2\|x_1 - x_*\|^2, \forall t \leq \tau\right\}$ happens where x_* is the optimal point in the domain. Though the idea is similar from an abstract level, things that need to be proved are very different, which is because we are in a non-convex world. More precisely, we use the initial function value gap rather than the initial distance to the optimal point to bound other terms. An immediate corollary from E_τ is that $F(x_{t+1}) - F_* \leq 2(F(x_1) - F_*)$, which implies $\|\nabla F(x_t)\| \leq 2\sqrt{L(F(x_1) - F_*)}$ for every $t \leq \tau + 1$ in a high probability. Now the insight is that $\|\nabla F(x_t)\|$ admitting a uniform upper bound through all iterations is highly possible. Thus, we can use this potential bound to clip the heavy-tailed stochastic gradient and drop the additional bounded pth moment estimates assumption. In comparison, the proof strategy in Cutkosky and Mehta (2021) is much different, in which the authors simply assume $\mathbb{E}[\|\widehat{\nabla} F(x)\|^p | x]$ is uniformly upper bounded. We note that this stronger assumption immediately implies that $\|\nabla F(x)\|$ has a uniform upper bound. However, as described above, the event E_τ is enough to help us choose the proper clipping magnitude. Hence, we can drop the extra assumption used in Cutkosky and Mehta (2021).

With the above plan, we can start the proof. We first introduce some notations used in the analysis. Let $\mathcal{F}_t$ be the natural filtration for both algorithms. Under this definition, x_t is $\mathcal{F}_{t-1}$

measurable. $\mathbb{E}_t[\cdot]$ is used to denote $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ for brevity. We also employ the following definitions:

$$\begin{aligned}
&\Delta_t = F(x_t) - F_*;\ \ \nabla_t = \nabla F(x_t);\\
&\epsilon_t = g_t - \nabla_t;\ \ \epsilon_t^u = g_t - \mathbb{E}_t[g_t];\ \ \epsilon_t^b = \mathbb{E}_t[g_t] - \nabla_t;\\
&\xi_t = \begin{cases} -\nabla_1 & t = 0 \\ d_t - \nabla_t & t \geq 1 \end{cases};\\
&Z_t = \begin{cases} \mathbb{1}_{t\geq 2}(\nabla_{t-1} - \nabla_t) & \text{For Algorithm 1} \\ \mathbb{1}_{t\geq 2}(\nabla f(x_t, \Xi_t) - \nabla f(x_{t-1}, \Xi_t) + \nabla_{t-1} - \nabla_t) & \text{For Algorithm 2} \end{cases}.
\end{aligned}$$

The decomposition of $\epsilon_t = \epsilon_t^u + \epsilon_t^b$ is proposed in Gorbunov et al. (2020). We borrow it here and first provide the bounds for $\|\epsilon_t^u\|$ and $\|\epsilon_t^b\|$ in Lemma 5

Lemma 5 *For both Algorithms 1 and 2, $\forall t \in [T]$, we have $\|\epsilon_t^u\| \leq 2M$. Besides, if $\|\nabla_t\| \leq M/2$, then there is*

$$\|\epsilon_t^b\| \leq 2\sigma^p M^{1-p};\ \ \mathbb{E}_t[\|\epsilon_t^u\|^2] \leq 10\sigma^p M^{2-p}.$$

Remark 6 *Similar results appear in Zhang et al. (2020) (for heavy-tailed noise) and Gorbunov et al. (2020) (for bounded variance) before, but the constants in ours are mildly tighter.*

The coefficient $1/2$ in the condition $\|\nabla_t\| \leq M/2$ simply follows the same choice as prior works. It can be changed to any number in $(0,1)$ with no essential difference. Recall that when E_τ holds, $\|\nabla_t\| \leq 2\sqrt{L\Delta_1}$ for any $t \leq \tau + 1$. To satisfy $\|\nabla_t\| \leq M/2$, naturally, M can be chosen larger than $4\sqrt{L\Delta_1}$. This answers why $4\sqrt{L\Delta_1}$ shows up in the choice of M for both Theorems 3 and 4.

Our next task is to express ξ_t via ξ_0, $Z_{s\leq t}$ and $\epsilon_{s\leq t}$ as shown in Lemma 7. This representation is common in the related literature and can help us to measure how fast the difference between d_t and ∇_t can decrease.

Lemma 7 *For both Algorithms 1 and 2, $\forall t \in [T]$, we have*

$$\xi_t = \beta^t \xi_0 + \beta\left(\sum_{s=1}^{t} \beta^{t-s} Z_s\right) + (1-\beta)\left(\sum_{s=1}^{t} \beta^{t-s}\epsilon_s\right).$$

Equipped with Lemma 7, the term $\eta\sum_{t=1}^{\tau}\|\nabla_t\| + \Delta_{\tau+1}$ in (6) can be upper bounded as follows.

Lemma 8 *For both Algorithms 1 and 2, $\forall \tau \in \{0\} \cup [T]$, we have*

$$\eta\sum_{t=1}^{\tau}\|\nabla_t\| + \Delta_{\tau+1} \leq \Delta_1 + \frac{\tau\eta^2 L}{2} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta}\mathbb{1}_{\tau\geq 1} + 2\eta\sum_{t=1}^{\tau}\beta\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\| + (1-\beta)\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\|.$$

Lemma 8 gives us some hints about the next step. For the term $\frac{\tau\eta^2 L}{2} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta}\mathbb{1}_{\tau\geq 1}$, by carefully choosing η and β, it can be bounded by $O(\Delta_1)$. Thus, we only need to care about $\|\sum_{s=1}^{t}\beta^{t-s}Z_s\|$ and $\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\|$. However, when Algorithm 2 is considered, the term $\|\sum_{s=1}^{t}\beta^{t-s}Z_s\|$ is essentially different from it in Cutkosky and Mehta (2021) since Z_s integrates the variance reduction part. Therefore, we need to come up with a way to quantify the effect of variance reduction in a high-probability style as noted before. Besides, the high-probability bound of $\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\|$ in

Cutkosky and Mehta (2021) can not be applied either due to no assumption of the bounded pth moment gradient estimates. So the departure from the existing works starts from here.

We first bound the easy one $\|\sum_{s=1}^{t}\beta^{t-s}Z_s\|$ in Lemma 9. As one can see, the bound for Algorithm 2 is roughly in the order of $\frac{\eta L}{\sqrt{1-\beta}}$. The acceleration from variance reduction is achieved by improving a factor of $\frac{1}{\sqrt{1-\beta}}$ compared with $\frac{\eta L}{1-\beta}$ for Algorithm 1. We also want to mention that assuming the almost surely smoothness for P2 (Assumption (4)) is critical in proving this variance-reduced high-probability bound. We refer the reader to the proof in Section B for more details.

Lemma 9 *We have that*

- *for Algorithm 1, $\forall t\in[T]$, we have* $\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\|\le\frac{\eta L}{1-\beta}$.
- *for Algorithm 2, $\forall t\in[T]$, we have* $\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\|\le\frac{9\eta L}{\sqrt{1-\beta}}\log\frac{3T}{\delta}$ *with probability at least* $1-\frac{\delta}{T}$.

Now we focus on the other term $\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\|$ and provide its first bound in Lemma 10.

Lemma 10 *For both Algorithms 1 and 2, $\forall t\in[T]$, we have*

$$\left\|\sum_{s=1}^{t}\beta_1^{t-s}\epsilon_s\right\|\le\left|\sum_{s=1}^{t}U_s^t\right|+\sqrt{2\left|\sum_{s=1}^{t}R_s^t\right|}+\sqrt{2\sum_{s=1}^{t}\mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right]}+\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s^b\right\|.$$

where U_s^t is a martingale difference sequence satisfying $\left|U_s^t\right|\le\left\|\beta^{t-s}\epsilon_s^u\right\|$ and $R_s^t=\left\|\beta^{t-s}\epsilon_s^u\right\|^2-\mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right]$ is also a martingale difference sequence.

We explain more here to help the reader to understand this complicated result better. The first step is to use $\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\|\le\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s^u\|+\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\|$, which is very intuitive since we want to use the bounds on $\mathbb{E}_s[\|\epsilon_s^u\|^2]$ and $\|\epsilon_s^b\|$ shown in Lemma 5. Next, we invoke a technical tool (Lemma 15 in Section A) to get $\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s^u\|\le\left|\sum_{s=1}^{t}U_s^t\right|+\sqrt{2\sum_{s=1}^{t}\|\beta^{t-s}\epsilon_s^u\|^2}$ where the definition of U_s^t is given in the proof. Finally, to let $\mathbb{E}_s[\|\epsilon_s^u\|^2]$ appear, we employ R_s^t to obtain the desired result. More details can be found in the proof in Section B.

Thus, our final task is to bound $\left|\sum_{s=1}^{t}U_s^t\right|$ and $\left|\sum_{s=1}^{t}R_s^t\right|$. As stated in Lemma 10, both of them are martingale difference sequences. Therefore we would like to use some concentration inequality. By observing that both U_s^t and R_s^t are bounded almost surely because of $\|\epsilon_s^u\|\le 2M$, we can apply the famous Bernstein Inequality (Lemma 13 in the appendix) and obtain that the following two events (Lemmas 11 and 12) happen with high probability.

Lemma 11 *For both Algorithms 1 and 2, $\forall t\in[T]$, we have $\Pr\left[a_t\right]\ge 1-\frac{\delta}{2T}$ where*

$$a_t=\left\{\left|\sum_{s=1}^{t}U_s^t\right|\le\left(\frac{4}{3}+2\sqrt{\frac{5\left(\sigma/M\right)^p}{1-\beta}}\right)M\log\frac{4T}{\delta}\ \textit{or}\ \sum_{s=1}^{t}\mathbb{E}_s\left[\left(U_s^t\right)^2\right]>\frac{10\sigma^pM^{2-p}}{1-\beta}\log\frac{4T}{\delta}\right\}.$$

Lemma 12 *For both Algorithms 1 and 2, $\forall t\in[T]$, we have $\Pr\left[b_t\right]\ge 1-\frac{\delta}{2T}$ where*

$$b_t=\left\{\left|\sum_{s=1}^{t}R_s^t\right|\le\left(\frac{16}{3}+4\sqrt{\frac{5\left(\sigma/M\right)^p}{1-\beta}}\right)M^2\log\frac{4T}{\delta}\ \textit{or}\ \sum_{s=1}^{t}\mathbb{E}_s\left[\left(R_s^t\right)^2\right]>\frac{40\sigma^pM^{4-p}}{1-\beta}\log\frac{4T}{\delta}\right\}.$$

We remark that the terms $\frac{10\sigma^p M^{2-p}}{1-\beta}\log\frac{4T}{\delta}$ in Lemma 11 and $\frac{40\sigma^p M^{4-p}}{1-\beta}\log\frac{4T}{\delta}$ in 12 are chosen carefully to finally let the inequalities on the conditional variance fail. Then these two events will degenerate to the bounds on $\left|\sum_{s=1}^t U_s^t\right|$ and $\left|\sum_{s=1}^t R_s^t\right|$, which are exactly what we need.

With the above lemmas, we are finally able to prove Theorem 4.

Proof [of Theorem 4] We will use induction to prove that the event $G_\tau = E_\tau \cap A_\tau \cap B_\tau \cap C_\tau$ holds with probability at least $1-\frac{2\tau\delta}{T}$ for any $\tau \in \{0\}\cup[T]$ where

$$E_\tau = \left\{\eta\sum_{s=1}^t \|\nabla_s\| + \Delta_{t+1} \le 2\Delta_1, \forall t \le \tau\right\};\;\; A_\tau = \cap_{t=1}^\tau a_t;\;\; B_\tau = \cap_{t=1}^\tau b_t;\;\; C_\tau = \cap_{t=1}^\tau c_t.$$

Events a_t and b_t are defined in Lemmas 11 and 12 respectively. c_t is from Lemma 9 defined as

$$c_t = \left\{\left\|\sum_{s=1}^t \beta^{t-s} Z_s\right\| \le \frac{9\eta L\log\frac{3T}{\delta}}{\sqrt{1-\beta}}\right\}. \tag{7}$$

Note that $E_0 = \{\Delta_1 \le 2\Delta_1\}$ can be viewed as the whole probability space. Thus, we have $A_0 = B_0 = C_0 = E_0$. Another useful fact is that $A_\tau = A_{\tau-1}\cap a_\tau$, $B_\tau = B_{\tau-1}\cap b_\tau$ and $C_\tau = C_{\tau-1}\cap c_\tau$.

Now we can start the induction. When $\tau=0$, $G_0 = E_0 = \{\Delta_1 \le 2\Delta_1\}$ holds with probability $1 = 1-\frac{2\tau\delta}{T}$. Next, suppose the induction hypothesis $\Pr[G_{\tau-1}] \ge 1-\frac{2(\tau-1)\delta}{T}$ is true for some $\tau\in[T]$. We will prove that $\Pr[G_\tau] \ge 1-\frac{2\tau\delta}{T}$. To start with, we consider the following event

$$E_{\tau-1}\cap A_\tau\cap B_\tau\cap C_\tau = G_{\tau-1}\cap a_\tau\cap b_\tau\cap c_\tau.$$

From Lemmas 11, 12 and 9, there are $\Pr[a_\tau] \ge 1-\frac{\delta}{2T}$, $\Pr[b_\tau]\ge 1-\frac{\delta}{2T}$ and $\Pr[c_\tau]\ge 1-\frac{\delta}{T}$. Combining with the induction hypothesis, we have $\Pr[E_{\tau-1}\cap A_\tau\cap B_\tau\cap C_\tau] \ge 1-\frac{2\tau\delta}{T}$.

Note that under the event $E_{\tau-1}$, there is $\Delta_t \le \eta\sum_{s=1}^{t-1}\|\nabla_s\| + \Delta_t \le 2\Delta_1$, $\forall t\le\tau$ which implies $\|\nabla_t\| \overset{(a)}{\le} \sqrt{2L\Delta_t} \le 2\sqrt{L\Delta_1} \overset{(b)}{\le} \frac{M}{2}$, $\forall t \le \tau$ where (a) is by Fact 2 and for (b) we use $M \ge 4\sqrt{L\Delta_1}$. Thus, in addition to $\|\epsilon_t^u\| \le 2M$, the following two bounds hold for any $t\le\tau$ by Lemma 5:

$$\mathbb{E}_t[\|\epsilon_t^u\|^2] \le 10\sigma^p M^{2-p};\;\; \|\epsilon_t^b\| \le 2\sigma^p M^{1-p}. \tag{8}$$

Equipped with (8), we can find that for any $t\le\tau$

$$\sum_{s=1}^t \mathbb{E}_s\left[\left(U_s^t\right)^2\right] \le \sum_{s=1}^t \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right] \le \sum_{s=1}^t \beta^{2t-2s}\cdot 10\sigma^p M^{2-p} \le \frac{10\sigma^p M^{2-p}}{1-\beta}; \tag{9}$$

$$\sum_{s=1}^t \mathbb{E}_s\left[\left(R_s^t\right)^2\right] \le \sum_{s=1}^t \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^4\right] \le \sum_{s=1}^t \beta^{4t-4s}\cdot 4M^2\cdot 10\sigma^p M^{2-p} \le \frac{40\sigma^p M^{4-p}}{1-\beta}. \tag{10}$$

Combining with the definitions of a_t and b_t, (9) and (10) imply that under the event $E_{\tau-1}\cap A_\tau\cap B_\tau\cap C_\tau$, the following two bounds hold for any $t\le\tau$,

$$\left|\sum_{s=1}^t U_s^t\right| \le \left(\frac{4}{3}+2\sqrt{\frac{5(\sigma/M)^p}{1-\beta}}\right)M\log\frac{4T}{\delta};\;\; \left|\sum_{s=1}^t R_s^t\right| \le \left(\frac{16}{3}+4\sqrt{\frac{5(\sigma/M)^p}{1-\beta}}\right)M^2\log\frac{4T}{\delta}. \tag{11}$$

Assuming $E_{\tau-1} \cap A_\tau \cap B_\tau \cap C_\tau$ happens, we invoke Lemma 8 for time τ to get

$$\begin{aligned}
\eta \sum_{t=1}^{\tau} \|\nabla_t\| + \Delta_{\tau+1} \leq & \Delta_1 + \frac{\tau\eta^2 L}{2} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + 2\eta \sum_{t=1}^{\tau} \beta \left\| \sum_{s=1}^{t} \beta^{t-s} Z_s \right\| + (1-\beta) \left\| \sum_{s=1}^{t} \beta^{t-s} \epsilon_s \right\| \\
\overset{(c)}{\leq} & \Delta_1 + \frac{\tau\eta^2 L}{2} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + \frac{18\beta\tau\eta^2 L \log\frac{3T}{\delta}}{\sqrt{1-\beta}} + 2\eta(1-\beta) \sum_{t=1}^{\tau} \left\| \sum_{s=1}^{t} \beta^{t-s} \epsilon_s \right\| \\
\leq & \Delta_1 + \frac{20\tau\eta^2 L \log\frac{4T}{\delta}}{\sqrt{1-\beta}} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + 2\eta(1-\beta) \sum_{t=1}^{\tau} \left\| \sum_{s=1}^{t} \beta^{t-s} \epsilon_s \right\| \\
\overset{(d)}{\leq} & \Delta_1 + \frac{20\tau\eta^2 L \log\frac{4T}{\delta}}{\sqrt{1-\beta}} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + 40\tau\eta M (1-\beta) \log\frac{4T}{\delta} \\
\overset{(e)}{\leq} & \Delta_1 + \frac{\Delta_1}{3} + \frac{\Delta_1}{3} + \frac{\Delta_1}{3} = 2\Delta_1
\end{aligned} \tag{12}$$

where in (c) we use the event c_t (see (7)) to bound $\| \sum_{s=1}^{t} \beta^{t-s} Z_s \|$. In (e), we plug in the choice of $\eta = \sqrt{\frac{\sqrt{1-\beta}\Delta_1}{60TL\log\frac{4T}{\delta}}} \wedge \frac{1-\beta}{9\beta}\sqrt{\frac{\Delta_1}{L}} \wedge \frac{\Delta_1}{120TM(1-\beta)\log\frac{4T}{\delta}}$. The term $\| \sum_{s=1}^{t} \beta^{t-s} \epsilon_s \|$ in (d) is bounded by first employing Lemma 10 to get

$$\begin{aligned}
\left\| \sum_{s=1}^{t} \beta^{t-s} \epsilon_s \right\| \leq & \left| \sum_{s=1}^{t} U_s^t \right| + \sqrt{2\left| \sum_{s=1}^{t} R_s^t \right|} + \sqrt{2 \sum_{s=1}^{t} \mathbb{E}_s \left[\|\beta^{t-s} \epsilon_s^u\|^2 \right]} + \left\| \sum_{s=1}^{t} \beta^{t-s} \epsilon_s^b \right\| \\
\overset{(f)}{\leq} & M \log\frac{4T}{\delta} \left(\frac{4}{3} + \sqrt{\frac{32}{3}} + \frac{2(\sigma/M)^p}{1-\beta} + 4\sqrt{5\frac{(\sigma/M)^p}{1-\beta}} + 2\sqrt{\sqrt{20\frac{(\sigma/M)^p}{1-\beta}}} \right) \\
\overset{(g)}{\leq} & 20M \log\frac{4T}{\delta}
\end{aligned}$$

where we use (8) and (11) in (f); (g) is due to $\frac{(\sigma/M)^p}{1-\beta} \leq 1$ by the choice of M. Now note that (12) implies $e_\tau = \{\eta \sum_{t=1}^{\tau} \|\nabla_t\| + \Delta_{\tau+1} \leq 2\Delta_1\}$ is a superset of $E_{\tau-1} \cap A_\tau \cap B_\tau \cap C_\tau$. Therefore

$$\Pr[G_\tau] = \Pr[E_{\tau-1} \cap e_\tau \cap A_\tau \cap B_\tau \cap C_\tau] = \Pr[E_{\tau-1} \cap A_\tau \cap B_\tau \cap C_\tau] \geq 1 - \frac{2\tau\delta}{T}.$$

Hence, the induction is completed.

Finally, we know $\Pr[E_T] \geq \Pr[G_T] \geq 1 - 2\delta$ which implies with probability at least $1 - 2\delta$, $\eta \sum_{t=1}^{T} \|\nabla_t\| + \Delta_{T+1} \leq 2\Delta_1$. By plugging our choices of β, M and η, we conclude

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla_t\| \leq \frac{2\Delta_1}{\eta T} = O\left(\frac{\sqrt{L\Delta_1 \log\frac{T}{\delta}}}{T^{\frac{3p-2}{4(2p-1)}}} \vee \frac{\sqrt{L\Delta_1}}{T^{\frac{p-1}{2p-1}}} \vee \frac{\sigma \log\frac{T}{\delta}}{T^{\frac{p-1}{2p-1}}} \vee \frac{\sqrt{L\Delta_1}\log\frac{T}{\delta}}{T^{\frac{p}{2p-1}}} \right).$$

■

5. Open Questions

It still remains some limitations in our work and there are many open problems worth exploring. First of all, we wonder whether our accelerated rate $O(\log(T/\delta)T^{\frac{1-p}{2p-1}})$ can be improved further or not when $F(x) = \mathbb{E}[f(x, \Xi)]$. We only know that the rate will reduce to $O(\log(T/\delta)T^{-1/3})$ when $p = 2$ matching the in-expectation lower bound $\Omega(T^{-1/3})$ up to a logarithmic factor. However, there is nothing known to us for $p \in (1, 2)$. Second, our accelerated result is proved under the assumption that $f(x, \Xi)$ is smooth almost surely. Whereas, we guess it is possible to relax it to the averaged smooth assumption, i.e., $\mathbb{E}_{\Xi\sim\mathcal{D}}[\|f(x, \Xi)-f(y, \Xi)\|^2] \leq L^2\|x-y\|^2$, which is the standard assumption used in Arjevani et al. (2019) for proving the lower bound when $p = 2$. Besides, as mentioned before, our analysis is not adaptive to the noise level σ. In other words, when $\sigma = 0$, our rate can not recover the well-known and optimal rate $\Theta(T^{-1/2})$ for deterministic algorithms. Thus, we believe it is an interesting task to improve our analysis in a further step. Additionally, it is still unclear how to remove the extra term T appearing in $\log(T/\delta)$. Finally, our choices of parameters heavily rely on the prior knowledge of the problem itself, which may be hard to know or even estimate in practice. Hence, it is important and worthful to find parameter-free algorithms that can achieve the same convergence rate for both two problems considered in this paper. We leave these questions as the future direction and look forward to them being addressed.

Acknowledgments

This work is generously supported by the National Science Foundation under the grant CCF-2106508. Additionally, Zhengyuan Zhou would like to acknowledge New York University's Center for Global Economy and Business faculty research grant during the 2023 – 2024 year. We are also grateful to the anonymous reviewers for their constructive comments and suggestions.

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Recapp: Crafting a more efficient catalyst for convex optimization. In *International Conference on Machine Learning*, pages 2658–2685. PMLR, 2022.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

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

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *arXiv preprint arXiv:1905.12412*, 2019.

Kfir Levy, Ali Kavis, and Volkan Cevher. Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

Deyi Liu, Lam M Nguyen, and Quoc Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. Adaptive accelerated (extra-) gradient methods with variance reduction. In *International Conference on Machine Learning*, pages 13947–13994. PMLR, 2022.

Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791. PMLR, 2013.

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

Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *arXiv preprint arXiv:1202.6258*, 2012.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 2013.

Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

Chaobing Song, Yong Jiang, and Yi Ma. Variance reduction via accelerated dual averaging for finite-sum optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Kaiwen Zhou, Qinghua Ding, Fanhua Shang, James Cheng, Danli Li, and Zhi-Quan Luo. Direct acceleration of saga using sampled negative momentum. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1602–1610. PMLR, 2019.

Appendix A. Technical Tools

In this section, we list some helpful technical results that appeared in the previous research, some proof of which will be omitted. The interested reader can refer to the original work for details.

The first inequality we need is the famous Bernstein Inequality for martingale difference sequence.

Lemma 13 *(Bernstein Inequality for martingale difference sequence in Bennett (1962); Dzhaparidze and Van Zanten (2001)) Suppose $X_{t\in[T]} \in \mathbb{R}$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_{t\in[T]}$ satisfying $|X_t| \le R$ almost surely for some constant R. Let $\sigma_t^2 = \mathbb{E}\left[|X_t|^2 \mid \mathcal{F}_{t-1}\right]$, then for any $a > 0$ and $F > 0$, there is*

$$\Pr\left[\left|\sum_{t=1}^{T} X_t\right| > a \text{ and } \sum_{t=1}^{T} \sigma_t^2 \le F\right] \le 2\exp\left(-\frac{a^2}{2F + 2Ra/3}\right).$$

The following concentration inequality proved in the general Hilbert Space is also useful. A similar result has appeared in Lemma 12 in Cutkosky and Mehta (2021). However, the term $\sum_{s=1}^{T} \sigma_s^2$ is stated as $\sum_{s=1}^{t} \sigma_s^2$ instead, which is not correct after private communication with the authors of Cutkosky and Mehta (2021). Therefore, we provide the correct version of this dimension-free concentration inequality here[2] with its proof.

Lemma 14 *(Corrected version of Lemma 12 in Cutkosky and Mehta (2021)) Suppose $X_{t\in[T]}$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_{t\in[T]}$ in a Hilbert Space satisfying $\|X_t\| \le R$ almost surely for some constant R and $\mathbb{E}\left[\|X_t\|^2 \mid \mathcal{F}_{t-1}\right] \le \sigma_t^2$ almost surely for some constant σ_t^2 . Then with probability at least $1-\delta$, $\forall t \in [T]$, there is*

$$\left\|\sum_{s=1}^{t} X_s\right\| \le 3R\log\frac{3}{\delta} + 3\sqrt{\sum_{s=1}^{T} \sigma_s^2 \log\frac{3}{\delta}}.$$

Proof By Lemma 15 (see below), for any $t \in [T]$ we have

$$\left\|\sum_{s=1}^{t} X_s\right\| \le \left|\sum_{s=1}^{t} M_s\right| + \sqrt{\max_{s\in[t]} \|X_s\|^2 + \sum_{s=1}^{t} \|X_s\|^2}$$

where $M_t \in \mathbb{R}$ is a martingale difference sequence satisfying $|M_t| \le \|X_t\|$ almost surely.

By using $\|X_t\| \le R$ almost surely, we have

$$\begin{aligned}
\left\|\sum_{s=1}^{t} X_s\right\| &\le \left|\sum_{s=1}^{t} M_s\right| + \sqrt{R^2 + \sum_{s=1}^{t} \|X_s\|^2} \\
&= \left|\sum_{s=1}^{t} M_s\right| + \sqrt{R^2 + \sum_{s=1}^{t} \underbrace{\|X_s\|^2 - \mathbb{E}\left[\|X_s\|^2 \mid \mathcal{F}_{s-1}\right]}_{U_s} + \sum_{s=1}^{t} \mathbb{E}\left[\|X_s\|^2 \mid \mathcal{F}_{s-1}\right]} \\
&\le \left|\sum_{s=1}^{t} M_s\right| + \sqrt{R^2 + \sum_{s=1}^{t} U_s + \sum_{s=1}^{t} \sigma_s^2}.
\end{aligned}$$

2. The results in Cutkosky and Mehta (2021) still hold after this correction.

where the last inequality is due to $\mathbb{E}\left[\|X_s\|^2 \mid \mathcal{F}_{s-1}\right] \le \sigma_s^2$ almost surely.

Note that

$$|M_t| \le R, \mathbb{E}\left[|M_t|^2 \mid \mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\|X_t\|^2 \mid \mathcal{F}_{t-1}\right] \le \sigma_t^2,$$

by Freedman's inequality Freedman (1975), there is

$$\Pr\left[\forall t \in [T], \left|\sum_{s=1}^{t} M_s\right| \le \frac{2R}{3}\log\frac{1}{\delta} + \sqrt{2\sum_{s=1}^{T}\sigma_s^2 \log\frac{1}{\delta}}\right] \ge 1-2\delta.$$

Similarly, we have

$$\Pr\left[\forall t \in [T], \sum_{s=1}^{t} U_s \le \frac{2R^2}{3}\log\frac{1}{\delta} + \sqrt{2\sum_{s=1}^{T}\sigma_s^2 R^2 \log\frac{1}{\delta}}\right] \ge 1-\delta.$$

Hence, with probability at least $1-3\delta$, for any $t \in [T]$

$$\begin{aligned}\left\|\sum_{s=1}^{t} X_s\right\| \le & \frac{2R}{3}\log\frac{1}{\delta} + \sqrt{2\sum_{s=1}^{T}\sigma_s^2\log\frac{1}{\delta}} + \sqrt{R^2 + \frac{2R^2}{3}\log\frac{1}{\delta} + \sqrt{2\sum_{s=1}^{T}\sigma_s^2 R^2\log\frac{1}{\delta}} + \sum_{s=1}^{t}\sigma_s^2} \\ \le & 3R\max\left\{1, \log\frac{1}{\delta}\right\} + 3\sqrt{\sum_{s=1}^{T}\sigma_s^2 \max\left\{1, \log\frac{1}{\delta}\right\}}.\end{aligned}$$

By changing δ to $\delta/3$, the proof is finished. ■

The last important tool is also proved by Cutkosky and Mehta (2021), the original statement of which is for the Banach Space. We simplify the result since only $\mathbb{R}^d$ is considered in this paper.

Lemma 15 *(Lemma 10 in Cutkosky and Mehta (2021)) Suppose $X_{t\in[T]} \in \mathbb{R}^d$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_{t\in[T]}$. Consider the sequence of real numbers Y_t defined as*

$$Y_t = \begin{cases} 0 & t = 0 \\ sgn\left(\sum_{i=1}^{t-1} Y_i\right)\frac{\langle\sum_{i=1}^{t-1} X_i, X_t\rangle}{\|\sum_{i=1}^{t-1} X_i\|} & t \ne 0 \text{ and } \sum_{i=1}^{t-1} X_i \ne \mathbf{0} \\ 0 & t \ne 0 \text{ and } \sum_{i=1}^{t-1} X_i = \mathbf{0} \end{cases}.$$

Then $Y_{t\in[T]}$ is a also martingale difference sequence satisfying $|Y_t| \le \|X_t\|, \forall t \in [T]$ and

$$\left\|\sum_{t=1}^{T} X_t\right\| \le \left|\sum_{t=1}^{T} Y_t\right| + \sqrt{\max_{t\in[T]}\|X_t\|^2 + \sum_{t=1}^{T}\|X_t\|^2}.$$

Appendix B. Missing Proofs In Section 4

In this section, we provide the omitted poofs of lemmas stated in Section 4. We first help the reader to recall the notations used in the analysis:

$$\begin{aligned}
&\Delta_t = F(x_t) - F_*;\ \ \nabla_t = \nabla F(x_t);\ \ \mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_{t-1}\right];\\
&\epsilon_t = g_t - \nabla_t;\ \ \epsilon_t^u = g_t - \mathbb{E}_t\left[g_t\right];\ \ \epsilon_t^b = \mathbb{E}_t\left[g_t\right] - \nabla_t;\\
&\xi_t = \begin{cases} -\nabla_1 & t = 0 \\ d_t - \nabla_t & t \geq 1 \end{cases};\\
&Z_t = \begin{cases} \mathbb{1}_{t\geq 2}\left(\nabla_{t-1} - \nabla_t\right) & \text{For Algorithm 1} \\ \mathbb{1}_{t\geq 2}\left(\nabla f(x_t, \Xi_t) - \nabla f(x_{t-1}, \Xi_t) + \nabla_{t-1} - \nabla_t\right) & \text{For Algorithm 2} \end{cases}
\end{aligned}$$

where $\mathcal{F}_t$ is the natural filtration.

B.1. Proof of Lemma 5

Proof First, from the definition of ϵ_t^u, we have $\|\epsilon_t^u\| \leq \|g_t\| + \|\mathbb{E}_t\left[g_t\right]\| \leq \|g_t\| + \mathbb{E}_t\left[\|g_t\|\right] \leq 2M$. For the second and the third inequalities, we use Algorithm 1 as an example. The same proof can be applied to Algorithm 2 directly.

For the second one, we know

$$\begin{aligned}
\left\|\epsilon_t^b\right\| = \|\mathbb{E}_t\left[g_t\right] - \nabla_t\| &= \left\|\mathbb{E}_t\left[g_t - \widehat{\nabla} F(x_t)\right]\right\| \leq \mathbb{E}_t\left[\left\|g_t - \widehat{\nabla} F(x_t)\right\|\right]\\
&= \mathbb{E}_t\left[\left\|\left(\frac{M}{\left\|\widehat{\nabla} F(x_t)\right\|} - 1\right)\widehat{\nabla} F(x_t)\mathbb{1}_{\|\widehat{\nabla} F(x_t)\|\geq M}\right\|\right]\\
&= \mathbb{E}_t\left[\left(\left\|\widehat{\nabla} F(x_t)\right\| - M\right)\mathbb{1}_{\|\widehat{\nabla} F(x_t)\|\geq M}\right]\\
&\leq \mathbb{E}_t\left[\left(\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\| + \|\nabla_t\| - M\right)\mathbb{1}_{\|\widehat{\nabla} F(x_t)\|\geq M}\right]\\
&\leq \mathbb{E}_t\left[\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|\mathbb{1}_{\|\widehat{\nabla} F(x_t)\|\geq M}\right]
\end{aligned}$$

where we use $\|\nabla_t\| \leq M/2 \leq M$ in the last step. Note that $\left\|\widehat{\nabla} F(x_t)\right\| \geq M \Rightarrow \left\|\widehat{\nabla} F(x_t) - \nabla_t\right\| \geq M/2$ when $\|\nabla_t\| \leq M/2$. Hence, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|\mathbb{1}_{\|\widehat{\nabla} F(x_t)\|\geq M}\right]\\
\leq&\mathbb{E}_t\left[\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|\mathbb{1}_{\|\widehat{\nabla} F(x_t) - \nabla_t\|\geq M/2}\right]\\
\overset{(a)}{\leq}&\mathbb{E}_t\left[\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|^p\right]^{1/p}\mathbb{E}_t\left[\mathbb{1}_{\|\widehat{\nabla} F(x_t) - \nabla_t\|\geq M/2}\right]^{1-1/p}\\
\leq&\sigma\Pr\left[\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|^p \geq (M/2)^p\right]^{1-1/p}\\
\overset{(b)}{\leq}&\sigma\left(\frac{2^p\sigma^p}{M^p}\right)^{1-1/p} = 2^{p-1}\sigma^p M^{1-p} \leq 2\sigma^p M^{1-p}
\end{aligned}$$

where (a) is due to Holder Inequality and (b) is because of Markov's Inequaliy.

For the third inequlaity, we prove it as follows

$$
\begin{aligned}
\mathbb{E}_t\left[\|\epsilon_t^u\|^2\right] &= \mathbb{E}_t\left[\|g_t - \mathbb{E}_t\left[g_t\right]\|^2\right] \overset{(c)}{\leq} \mathbb{E}_t\left[\|g_t - \nabla_t\|^2\right] \\
&= \mathbb{E}_t\left[\|g_t - \nabla_t\|^2 \mathbb{1}_{\|\widehat{\nabla} F(x_t)\| \geq M} + \|g_t - \nabla_t\|^2 \mathbb{1}_{\|\widehat{\nabla} F(x_t)\| < M}\right] \\
&= \mathbb{E}_t\left[\left\|\frac{M}{\left\|\widehat{\nabla} F(x_t)\right\|}\widehat{\nabla} F(x_t) - \nabla_t\right\|^2 \mathbb{1}_{\|\widehat{\nabla} F(x_t)\| \geq M} + \left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|^2 \mathbb{1}_{\|\widehat{\nabla} F(x_t)\| < M}\right] \\
&\overset{(d)}{\leq} \mathbb{E}_t\left[\frac{9}{4}M^2 \mathbb{1}_{\|\widehat{\nabla} F(x_t)\| \geq M} + \left(\frac{3}{2}M\right)^{2-p}\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\|^p\right] \\
&\leq \frac{9}{4}M^2 \frac{2^p \sigma^p}{M^p} + \left(\frac{3}{2}M\right)^{2-p}\sigma^p = \left[9 \cdot 2^{p-2} + \left(\frac{3}{2}\right)^{2-p}\right]\sigma^p M^{2-p} \\
&\leq 10\sigma^p M^{2-p}.
\end{aligned}
$$

where (c) is due to $\mathbb{E}_t\left[\|g_t - \mathbb{E}_t\left[g_t\right]\|^2\right] \leq \mathbb{E}_t\left[\|g_t - Y\|^2\right]$ for any $Y \in \mathcal{F}_{t-1}$. (d) is by when $\|\nabla_t\| \leq M/2$ there are

$$
\left\|\frac{M}{\left\|\widehat{\nabla} F(x_t)\right\|}\widehat{\nabla} F(x_t) - \nabla_t\right\| \leq M + \|\nabla_t\| \leq 3M/2
$$

and

$$
\left\|\widehat{\nabla} F(x_t) - \nabla_t\right\| \mathbb{1}_{\|\widehat{\nabla} F(x_t)\| < M} \leq M + \|\nabla_t\| \leq 3M/2.
$$

■

B.2. Proof of Lemma 7

Proof We first check for Algorithm 1. Use the definition of ϵ_t, Z_t and ξ_t here to get for $t \geq 2$

$$
\xi_t = d_t - \nabla_t = \beta d_{t-1} + (1-\beta)\, g_t - \nabla_t = \beta \xi_{t-1} + \beta Z_t + (1-\beta)\, \epsilon_t.
$$

Note that the above equation also holds when $t = 1$. Next, we calculate ξ_t for Algorithm 2 by noticing for $t \geq 2$

$$
\begin{aligned}
\xi_t &= d_t - \nabla_t = \beta d_{t-1} + (1-\beta)\, g_t + \beta\left(\nabla f(x_t, \Xi_t) - \nabla f(x_{t-1}, \Xi_t)\right) - \nabla_t \\
&= \beta \xi_{t-1} + \beta Z_t + (1-\beta)\, \epsilon_t.
\end{aligned}
$$

This equation is true for $t = 1$ again. We use this recursion for all iterations to finish the proof. ■

B.3. Proof of Lemma 8

Proof We will first prove $\Delta_{t+1}-\Delta_t\leq-\eta\left\|\nabla_t\right\|+2\eta\left\|\xi_t\right\|+\frac{\eta^2L}{2}$. This result has been shown in Cutkosky and Mehta (2021). We provide the analysis below for completeness. Starting with Fact 1

$$\begin{aligned}\Delta_{t+1}-\Delta_t=F(x_{t+1})-F(x_t)&\leq\langle\nabla_t,x_{t+1}-x_t\rangle+\frac{L}{2}\left\|x_{t+1}-x_t\right\|^2\\&=-\eta\langle\nabla_t,\frac{d_t}{\|d_t\|}\rangle+\frac{\eta^2L}{2}=-\eta\left\|d_t\right\|+\eta\langle\xi_t,\frac{d_t}{\|d_t\|}\rangle+\frac{\eta^2L}{2}\\&\overset{(a)}{\leq}-\eta\left\|d_t\right\|+\eta\left\|\xi_t\right\|+\frac{\eta^2L}{2}\overset{(b)}{\leq}-\eta\left\|\nabla_t\right\|+2\eta\left\|\xi_t\right\|+\frac{\eta^2L}{2}\end{aligned}$$

where (a) is by $\langle\xi_t,\frac{d_t}{\|d_t\|}\rangle\leq\left\|\xi_t\right\|\left\|\frac{d_t}{\|d_t\|}\right\|=\left\|\xi_t\right\|$ and (b) is due to $\left\|\nabla_t\right\|\leq\left\|d_t\right\|+\left\|\xi_t\right\|$. Now summing up from $t=1$ to τ, we obtain

$$\begin{aligned}&\eta\sum_{t=1}^{\tau}\left\|\nabla_t\right\|+\Delta_{\tau+1}\leq\Delta_1+\frac{\tau\eta^2L}{2}+2\eta\sum_{t=1}^{\tau}\left\|\xi_t\right\|\\&\overset{(c)}{=}\Delta_1+\frac{\tau\eta^2L}{2}+2\eta\sum_{t=1}^{\tau}\left\|\beta^t\xi_0+\beta\left(\sum_{s=1}^{t}\beta^{t-s}Z_s\right)+(1-\beta)\left(\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right)\right\|\\&\leq\Delta_1+\frac{\tau\eta^2L}{2}+2\eta\sum_{t=1}^{\tau}\beta^t\left\|\xi_0\right\|+\beta\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\|+(1-\beta)\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\|\\&\leq\Delta_1+\frac{\tau\eta^2L}{2}+\frac{2\beta\eta\left\|\xi_0\right\|}{1-\beta}\mathbb{1}_{\tau\geq1}+2\eta\sum_{t=1}^{\tau}\beta\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\|+(1-\beta)\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\|\\&\overset{(d)}{\leq}\Delta_1+\frac{\tau\eta^2L}{2}+\frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta}\mathbb{1}_{\tau\geq1}+2\eta\sum_{t=1}^{\tau}\beta\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\|+(1-\beta)\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\|\end{aligned}$$

where we invoke Lemma 7 in (c); Fact 2 leads to $2\left\|\xi_0\right\|=2\left\|\nabla_1\right\|\leq3\sqrt{L\Delta_1}$ in (d). ■

B.4. Proof of Lemma 9

Proof For Algorithm 1, from the definition of Z_s, we know

$$\left\|Z_s\right\|=\left\|\mathbb{1}_{s\geq2}\left(\nabla_{s-1}-\nabla_s\right)\right\|\leq\mathbb{1}_{s\geq2}\left\|\nabla_{s-1}-\nabla_s\right\|\leq\mathbb{1}_{s\geq2}L\left\|x_{s-1}-x_s\right\|\leq\eta L$$

where we use the smoothness assumption and note that $\left\|x_{s-1}-x_s\right\|=\left\|\eta\frac{d_s}{\|d_s\|}\right\|=\eta$. Hence,

$$\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\|\leq\sum_{s=1}^{t}\beta^{t-s}\left\|Z_s\right\|\leq\sum_{s=1}^{t}\beta^{t-s}\eta L\leq\frac{\eta L}{1-\beta}.$$

For Algorithm 2, let $t\in[T]$ be fixed and consider $s\in[t]$. From the definition of Z_s, we know $\beta^{t-s}Z_s$ is adapted to $\mathcal{F}_s$. Besides,

$$\mathbb{E}_s\left[\beta^{t-s}Z_s\right]=\beta^{t-s}\mathbb{E}_s\left[\mathbb{1}_{s\geq2}\left(\nabla f(x_s,\Xi_s)-\nabla f(x_{s-1},\Xi_s)+\nabla_{s-1}-\nabla_s\right)\right]=0$$

which implies $\beta^{t-s}Z_s$ is a martingale difference sequence. Note that we have $\left\|\beta^{t-1}Z_1\right\| = 0$ and for $s \geq 2$, by the almost surely smoothness assumption,

$$\left\|\beta^{t-s}Z_s\right\| \leq \left\|\nabla f(x_s, \Xi_s) - \nabla f(x_{s-1}, \Xi_s)\right\| + \left\|\nabla_s - \nabla_{s-1}\right\| \leq 2L\left\|x_s - x_{s-1}\right\| = 2\eta L.$$

Besides, there is

$$\begin{aligned}\mathbb{E}_s\left[\left\|\beta^{t-s}Z_s\right\|^2\right] &= \beta^{2t-2s}\mathbb{E}_s\left[\left\|\mathbb{1}_{s\geq 2}\left(\nabla f(x_s, \Xi_s) - \nabla f(x_{s-1}, \Xi_s) + \nabla_{s-1} - \nabla_s\right)\right\|^2\right] \\ &\leq \beta^{2t-2s}\mathbb{1}_{s\geq 2}\mathbb{E}_s\left[\left\|\nabla f(x_s, \Xi_s) - \nabla f(x_{s-1}, \Xi_s)\right\|^2\right] \leq \beta^{2t-2s}\eta^2 L^2.\end{aligned}$$

Now we invoke Lemma 14 to obtain that with probability at least $1-\delta$, for any $\tau \in [t]$, there is

$$\begin{aligned}\left\|\sum_{s=1}^{\tau}\beta^{t-s}Z_s\right\| &\leq 6\eta L\log\frac{3}{\delta} + 3\sqrt{\sum_{s=1}^{\tau}\beta^{2t-2s}\eta^2L^2\log\frac{3}{\delta}} \\ &\leq 3\eta L\left(2\log\frac{3}{\delta} + \sqrt{\frac{\log\frac{3}{\delta}}{1-\beta}}\right) \leq \frac{9\eta L\log\frac{3}{\delta}}{\sqrt{1-\beta}}.\end{aligned}$$

We choose $\tau = t$ and replace δ by $\frac{\delta}{T}$ to finish the proof. ■

B.5. Proof of Lemma 10

Proof Starting with the definition of $\epsilon_s = \epsilon_s^b + \epsilon_s^u$ to get

$$\left\|\sum_{s=1}^{t}\beta_1^{t-s}\epsilon_s\right\| \leq \left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s^u\right\| + \left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s^b\right\|.$$

Now we define the sequence U_s^t for $s \in \{0\} \cup [t]$

$$U_s^t = \begin{cases}\mathbf{0} & s = 0 \\ \operatorname{sgn}\left(\sum_{i=1}^{s-1}U_i^t\right)\frac{\left\langle\sum_{i=1}^{s-1}\beta^{t-i}\epsilon_i^u, \beta^{t-s}\epsilon_s^u\right\rangle}{\left\|\sum_{i=1}^{s-1}\beta^{t-i}\epsilon_i^u\right\|} & s \neq 0 \text{ and } \sum_{i=1}^{s-1}\beta^{t-i}\epsilon_i^u \neq \mathbf{0} \\ \mathbf{0} & s \neq 0 \text{ and } \sum_{i=1}^{s-1}\beta^{t-i}\epsilon_i^u = \mathbf{0}\end{cases}$$

According to Lemma 15, U_s^t is a martingale difference sequence satisfying $\left|U_s^t\right| \leq \left\|\beta^{t-s}\epsilon_s^u\right\|$ and

$$\begin{aligned}\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s^u\right\| &\leq \left|\sum_{s=1}^{t}U_s^t\right| + \sqrt{\max_{s\in[t]}\left\|\beta^{t-s}\epsilon_s^u\right\|^2 + \sum_{s=1}^{t}\left\|\beta^{t-s}\epsilon_s^u\right\|^2} \\ &\leq \left|\sum_{s=1}^{t}U_s^t\right| + \sqrt{2\sum_{s=1}^{t}\left\|\beta^{t-s}\epsilon_s^u\right\|^2 - \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right] + \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right]} \\ &\leq \left|\sum_{s=1}^{t}U_s^t\right| + \sqrt{2\left|\sum_{s=1}^{t}R_s^t\right|} + \sqrt{2\sum_{s=1}^{t}\mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right]}\end{aligned}$$

where $R_s^t = \left\|\beta^{t-s}\epsilon_s^u\right\|^2 - \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right]$. The proof is completed now. ■

B.6. Proof of Lemma 11

Proof For any fixed $t \in [T]$, note that U_s^t is a martingale difference sequence satisfying $\left|U_s^t\right| \leq \left\|\beta^{t-s}\epsilon_s^u\right\| \leq 2M$ (Lemmas 5 and 10). Hence by Bernstein Iinequality (Lemma 13), we know

$$\Pr\left[\left|\sum_{s=1}^{t} U_s^t\right| > a \text{ and } \sum_{s=1}^{t} \mathbb{E}_s\left[\left(U_s^t\right)^2\right] \leq F \log\frac{4T}{\delta}\right] \leq 2\exp\left(-\frac{a^2}{2F\log\frac{4T}{\delta} + 4Ma/3}\right).$$

Choose $a > 0$ here to satisfy

$$2\exp\left(-\frac{a^2}{2F\log\frac{4T}{\delta} + 4Ma/3}\right) = \frac{\delta}{2T} \Rightarrow a = \left(\frac{2}{3}M + \sqrt{\frac{4}{9}M^2 + 2F}\right)\log\frac{4T}{\delta}.$$

Next, we take $F = \frac{10\sigma^p M^{2-p}}{1-\beta}$. Thus, with probability at least $1 - \frac{\delta}{2T}$, there is

$$\left|\sum_{s=1}^{t} U_s^t\right| \leq \left(\frac{2}{3} + \sqrt{\frac{4}{9} + \frac{20\,(\sigma/M)^p}{1-\beta}}\right) M\log\frac{4T}{\delta} \leq \left(\frac{4}{3} + 2\sqrt{\frac{5\,(\sigma/M)^p}{1-\beta}}\right) M\log\frac{4T}{\delta}$$

or

$$\sum_{s=1}^{t} \mathbb{E}_s\left[\left(U_s^t\right)^2\right] > \frac{10\sigma^p M^{2-p}}{1-\beta}\log\frac{4T}{\delta}.$$

■

B.7. Proof of Lemma 12

Proof For any fixed $t \in [T]$, by Lemma 5, $R_s^t = \left\|\beta^{t-s}\epsilon_s^u\right\|^2 - \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right]$ is a martingale difference sequence satisfying

$$\left|R_s^t\right| \leq \left\|\beta^{t-s}\epsilon_s^u\right\|^2 + \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2\right] \leq 8M^2.$$

Hence by Bernstein inequality (Lemma 13), we know

$$\Pr\left[\left|\sum_{s=1}^{t} R_s^t\right| > a \text{ and } \sum_{s=1}^{t} \mathbb{E}_s\left[\left(R_s^t\right)^2\right] \leq F \log\frac{4T}{\delta}\right] \leq 2\exp\left(-\frac{a^2}{2F\log\frac{4T}{\delta} + 16M^2a/3}\right).$$

We choose $a > 0$ to satisfy

$$2\exp\left(-\frac{a^2}{2F\log\frac{4T}{\delta} + 16M^2a/3}\right) = \frac{\delta}{2T} \Rightarrow a = \left(\frac{8}{3}M^2 + \sqrt{\frac{64}{9}M^4 + 2F}\right)\log\frac{4T}{\delta}.$$

We take $F = \frac{40\sigma^p M^{4-p}}{1-\beta}$ here to obtain with probability at least $1 - \frac{\delta}{2T}$, there is

$$\left|\sum_{s=1}^{t} R_s^t\right| \leq \left(\frac{8}{3} + \sqrt{\frac{64}{9} + \frac{80\,(\sigma/M)^p}{1-\beta}}\right) M^2\log\frac{4T}{\delta} \leq \left(\frac{16}{3} + 4\sqrt{\frac{5\,(\sigma/M)^p}{1-\beta}}\right) M^2\log\frac{4T}{\delta}$$

or

$$\sum_{s=1}^{t} \mathbb{E}_s\left[\left(R_s^t\right)^2\right] > \frac{40\sigma^p M^{4-p}}{1-\beta}\log\frac{4T}{\delta}.$$

■

Appendix C. Proof of Theorem 3

The proof of Theorem 3 is almost the same as the proof of Theorem 4 under our unified analysis framework.

Proof We will use induction to prove that the event $G_\tau = E_\tau \cap A_\tau \cap B_\tau$ holds with probability at least $1 - \frac{\tau\delta}{T}$ for any $\tau \in \{0\} \cup [T]$ where

$$E_\tau = \left\{ \eta \sum_{s=1}^{t} \|\nabla_s\| + \Delta_{t+1} \le 2\Delta_1, \forall t \le \tau \right\}; \;\; A_\tau = \cap_{t=1}^{\tau} a_t; \;\; B_\tau = \cap_{t=1}^{\tau} b_t;$$

and

$$a_t = \left\{ \left| \sum_{s=1}^{t} U_s^t \right| \le \left(\frac{4}{3} + 2\sqrt{\frac{5\left(\sigma/M\right)^p}{1-\beta}} \right) M \log\frac{4T}{\delta} \text{ or } \sum_{s=1}^{t} \mathbb{E}_s\left[\left(U_s^t\right)^2 \right] > \frac{10\sigma^p M^{2-p}}{1-\beta} \log\frac{4T}{\delta} \right\};$$

$$b_t = \left\{ \left| \sum_{s=1}^{t} R_s^t \right| \le \left(\frac{16}{3} + 4\sqrt{\frac{5\left(\sigma/M\right)^p}{1-\beta}} \right) M^2 \log\frac{4T}{\delta} \text{ or } \sum_{s=1}^{t} \mathbb{E}_s\left[\left(R_s^t\right)^2 \right] > \frac{40\sigma^p M^{4-p}}{1-\beta} \log\frac{4T}{\delta} \right\}.$$

Note that $E_0 = \{\Delta_1 \le 2\Delta_1\}$ can be viewed as the whole probability space. Thus, we have $A_0 = B_0 = E_0$. Another useful fact is that $A_\tau = A_{\tau-1} \cap a_\tau$ and $B_\tau = B_{\tau-1} \cap b_\tau$.

Now we can start the induction. When $\tau = 0$, $G_0 = E_0 = \{\Delta_1 \le 2\Delta_1\}$ holds with probability $1 = 1 - \frac{\tau\delta}{T}$. Next, suppose the induction hypothesis $\Pr\left[G_{\tau-1}\right] \ge 1 - \frac{(\tau-1)\delta}{T}$ is true for some $\tau \in [T]$. We will prove that $\Pr\left[G_\tau\right] \ge 1 - \frac{\tau\delta}{T}$. To start with, we consider the following event

$$E_{\tau-1} \cap A_\tau \cap B_\tau = G_{\tau-1} \cap a_\tau \cap b_\tau.$$

From Lemmas 11 and 12, there are $\Pr\left[a_\tau\right] \ge 1 - \frac{\delta}{2T}$ and $\Pr\left[b_\tau\right] \ge 1 - \frac{\delta}{2T}$. Combining with the induction hypothesis, we obtain$\Pr\left[E_{\tau-1} \cap A_\tau \cap B_\tau\right] \ge 1 - \frac{\tau\delta}{T}$.

Note that under the event $E_{\tau-1}$, there is $\Delta_t \le \eta \sum_{s=1}^{t-1} \|\nabla_s\| + \Delta_t \le 2\Delta_1$, $\forall t \le \tau$ which implies $\|\nabla_t\| \overset{(a)}{\le} \sqrt{2L\Delta_t} \le 2\sqrt{L\Delta_1} \overset{(b)}{\le} \frac{M}{2}$, $\forall t \le \tau$ where (a) is by Fact 2 and for (b) we use $M \ge 4\sqrt{L\Delta_1}$. Thus, in addition to $\|\epsilon_t^u\| \le 2M$, the following two bounds hold for any $t \le \tau$ by Lemma 5:

$$\mathbb{E}_t\left[\|\epsilon_t^u\|^2 \right] \le 10\sigma^p M^{2-p}. \tag{13}$$

$$\left\|\epsilon_t^b\right\| \le 2\sigma^p M^{1-p}. \tag{14}$$

Equipped with (13) and (14), we can find that for any $t \le \tau$

$$\sum_{s=1}^{t} \mathbb{E}_s\left[\left(U_s^t\right)^2 \right] \le \sum_{s=1}^{t} \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^2 \right] \le \sum_{s=1}^{t} \beta^{2t-2s} \cdot 10\sigma^p M^{2-p} \le \frac{10\sigma^p M^{2-p}}{1-\beta}; \tag{15}$$

$$\sum_{s=1}^{t} \mathbb{E}_s\left[\left(R_s^t\right)^2 \right] \le \sum_{s=1}^{t} \mathbb{E}_s\left[\left\|\beta^{t-s}\epsilon_s^u\right\|^4 \right] \le \sum_{s=1}^{t} \beta^{4t-4s} \cdot 4M^2 \cdot 10\sigma^p M^{2-p} \le \frac{40\sigma^p M^{4-p}}{1-\beta}. \tag{16}$$

Combining with the definitions of events a_t and b_t, (15) and (16) imply that under the event $E_{\tau-1} \cap A_\tau \cap B_\tau$, the following two bounds hold for any $t \le \tau$,

$$\left|\sum_{s=1}^{t} U_s^t\right| \le \left(\frac{4}{3} + 2\sqrt{\frac{5\left(\sigma/M\right)^p}{1-\beta}}\right) M \log\frac{4T}{\delta}; \tag{17}$$

$$\left|\sum_{s=1}^{t} R_s^t\right| \le \left(\frac{16}{3} + 4\sqrt{\frac{5\left(\sigma/M\right)^p}{1-\beta}}\right) M^2 \log\frac{4T}{\delta}. \tag{18}$$

Assuming $E_{\tau-1} \cap A_\tau \cap B_\tau$ holds, we invoke Lemma 8 for time τ to get

$$\begin{aligned}
&\eta \sum_{t=1}^{\tau} \|\nabla_t\| + \Delta_{\tau+1} \\
\le&\Delta_1 + \frac{\tau\eta^2 L}{2} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + 2\eta\sum_{t=1}^{\tau}\beta\left\|\sum_{s=1}^{t}\beta^{t-s}Z_s\right\| + (1-\beta)\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\| \\
\overset{(c)}{\le}&\Delta_1 + \frac{\tau\eta^2 L}{2} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + \frac{2\beta\tau\eta^2 L}{1-\beta} + 2\eta\left(1-\beta\right)\sum_{t=1}^{\tau}\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\| \\
\le&\Delta_1 + \frac{2\tau\eta^2 L}{1-\beta} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + 2\eta\left(1-\beta\right)\sum_{t=1}^{\tau}\left\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\right\| \\
\overset{(d)}{\le}&\Delta_1 + \frac{2\tau\eta^2 L}{1-\beta} + \frac{3\beta\eta\sqrt{L\Delta_1}}{1-\beta} + 40\tau\eta M\left(1-\beta\right)\log\frac{4T}{\delta} \\
\overset{(e)}{\le}&\Delta_1 + \frac{\Delta_1}{3} + \frac{\Delta_1}{3} + \frac{\Delta_1}{3} \le 2\Delta_1
\end{aligned}$$

where (c) is due to Lemma 9. The bound of $\|\sum_{s=1}^{t}\beta^{t-s}\epsilon_s\|$ in (d) is totally the same as it in (12) in the proof of Theorem 4. We plug in the choice of $\eta = \sqrt{\frac{(1-\beta)\Delta_1}{6TL}} \wedge \frac{1-\beta}{9\beta}\sqrt{\frac{\Delta_1}{L}} \wedge \frac{\Delta_1}{120TM(1-\beta)\log\frac{4T}{\delta}}$ in (e). Now note this result implies the event $e_\tau = \{\eta\sum_{t=1}^{\tau}\|\nabla_t\| + \Delta_{\tau+1} \le 2\Delta_1\}$ is a superset of $E_{\tau-1} \cap A_\tau \cap B_\tau$. Therefore

$$\Pr\left[G_\tau\right] = \Pr\left[E_{\tau-1} \cap e_\tau \cap A_\tau \cap B_\tau\right] = \Pr\left[E_{\tau-1} \cap A_\tau \cap B_\tau\right] \ge 1 - \frac{\tau\delta}{T}.$$

Hence, the induction is completed.

Finally we know $\Pr\left[E_T\right] \geq \Pr\left[G_T\right] \geq 1-\delta$ which implies with probability at least $1-\delta$, $\eta \sum_{t=1}^{T}\|\nabla_t\| + \Delta_{T+1} \leq 2\Delta_1$. By plugging our choices of β, M and η, we conclude

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\|\nabla_t\| \leq \frac{2\Delta_1}{\eta T} \\
=& \frac{2\Delta_1}{\sqrt{\frac{T(1-\beta)\Delta_1}{6L}} \wedge \frac{T(1-\beta)}{9\beta}\sqrt{\frac{\Delta_1}{L}} \wedge \frac{\Delta_1}{120M(1-\beta)\log\frac{4T}{\delta}}} \\
=& O\left(\frac{\sqrt{L\Delta_1}}{\sqrt{T\left(1-\beta\right)}} \vee \frac{\beta\sqrt{L\Delta_1}}{T\left(1-\beta\right)} \vee M\log\frac{T}{\delta}\right) \\
=& O\left(\frac{\sqrt{L\Delta_1}}{\sqrt{T\left(1-\beta\right)}} \vee \frac{\beta\sqrt{L\Delta_1}}{T\left(1-\beta\right)} \vee \sigma\left(1-\beta\right)^{1-1/p}\log\frac{T}{\delta} \vee \sqrt{L\Delta_1}\left(1-\beta\right)\log\frac{T}{\delta}\right). \\
=& O\left(\frac{\sqrt{L\Delta_1}}{T^{\frac{p-1}{3p-2}}} \vee \frac{\sigma\log\frac{T}{\delta}}{T^{\frac{p-1}{3p-2}}} \vee \frac{\sqrt{L\Delta_1}\log\frac{T}{\delta}}{T^{\frac{p}{3p-2}}}\right).
\end{aligned}
$$

■