Online Convex Optimization with Heavy Tails: Old Algorithms, New Regrets, and Applications

Zijian Liu*

Abstract

In Online Convex Optimization (OCO), when the stochastic gradient has a finite variance, many algorithms provably work and guarantee a sublinear regret. However, limited results are known if the gradient estimate has a heavy tail, i.e., the stochastic gradient only admits a finite p-th central moment for some $\mathsf{p} \in (1, 2]$. Motivated by it, this work examines different old algorithms for OCO (e.g., Online Gradient Descent) in the more challenging heavy-tailed setting. Under the standard bounded domain assumption, we establish new regrets for these classical methods without any algorithmic modification. Remarkably, these regret bounds are fully optimal in all parameters (can be achieved even without knowing p), suggesting that OCO with heavy tails can be solved effectively without any extra operation (e.g., gradient clipping). Our new results have several applications. A particularly interesting one is the first provable convergence result for nonsmooth nonconvex optimization under heavy-tailed noise without gradient clipping. Furthermore, we explore broader settings (e.g., smooth OCO) and extend our ideas to optimistic algorithms to handle different cases simultaneously.

1 Introduction

This paper studies the online learning problem with convex losses, also known as Online Convex Optimization (OCO), a widely applicable framework that learns under streaming data [5, 11, 29, 38]. OCO has tons of implications for both designing and analyzing algorithms in different areas, for example, stochastic optimization [10, 25, 17], PAC learning [4], control theory [1, 12], etc.

In an OCO problem, a learning algorithm A would interact with the environment in T rounds, where $T \in \mathbb{N}$ can be either known or unknown. Formally, in each round round t, the learner A first decides an output $\boldsymbol{x}_t \in \mathcal{X}$ from a convex feasible set $\mathcal{X} \subseteq \mathbb{R}^d$, then the environment reveals a convex loss function $\ell_t : \mathcal{X} \to \mathbb{R}$, and A incurs a loss of $\ell_t(\boldsymbol{x}_t)$. After T many rounds, the quantity measuring the algorithm's performance is called regret, defined relative to any fixed competitor $\boldsymbol{x} \in \mathcal{X}$ as follows:

$$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}) \triangleq \sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}).$$

In the classical setting, instead of observing full information about ℓ_t, the learner A is only guaranteed to receive a subgradient $\nabla \ell_t(\boldsymbol{x}_t) \in \partial \ell_t(\boldsymbol{x}_t)$ at its decision, where $\partial \ell_t(\boldsymbol{x}_t)$ denotes the subdifferential set of ℓ_t at $\boldsymbol{x}_t$ [36]. This turns out to be enough for our purpose of minimizing the regret, since any OCO problem can be reduced to an Online Linear Optimization (OLO) instance via the inequality $\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}) \le \langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x} \rangle$, which holds due to convexity. Under the standard bounded domain assumption, i.e., $\mathcal{X}$ has a finite diameter D, many classical algorithms, e.g., Online Gradient Descent (OGD) [54], guarantee an optimal sublinear regret $GD\sqrt{T}$ for G-Lipschitz ℓ_t. Even better, in the case that computing an exact subgradient is intractable, and one could only query a stochastic estimate $\boldsymbol{g}_t$ satisfying $\mathbb{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] \in \partial \ell_t(\boldsymbol{x}_t)$, the OGD algorithm can still solve OCO effectively with a provable $(G+\sigma)D\sqrt{T}$ regret bound in expectation if the stochastic noise $\boldsymbol{g}_t - \nabla \ell_t(\boldsymbol{x}_t)$ has a bounded second moment σ^2 for some $\sigma \ge 0$, which is called the finite variance condition.

However, many works have pointed out that even for the easier stochastic optimization (i.e., $\ell_t = F$ for a common F), the typical finite variance assumption is too optimistic and can be violated in different tasks

*Stern School of Business, New York University, zl3067@stern.nyu.edu.

[13, 40, 49], and their observations suggest that the stochastic gradient only admits a finite p-th central moment upper bounded by σ^{p} for some $\mathsf{p} \in (1, 2]$, which is named heavy-tailed noise. This new assumption generalizes the classical finite variance condition ($\mathsf{p} = 2$) and becomes challenging when $\mathsf{p} < 2$. A particular evidence is that the famous Stochastic Gradient Descent (SGD) algorithm [35] (which is exactly OGD for stochastic optimization) provably diverges [49].

Though heavy-tailed stochastic optimization has been extensively studied [21, 28, 37], limited results are known for OCO with heavy tails. The only work under this topic that we are aware of is [51], which established a parameter-free regret bound in high probability (more discussions provided later). However, their algorithm includes many nontrivial modifications like gradient clipping and significantly deviates from the existing simple OCO algorithms used in practice. Especially, consider OGD as an example. Though the heavy-tailed issue is known, OGD (or just think of it as SGD) still works (sometimes very well) in practice even without gradient clipping and is arguably one of the most popular optimizers, which seemingly contradicts the theory of unconvergence mentioned before. This indicates that, for classical OCO algorithms under heavy-tailed noise, a huge gap exists between the empirical convergence (or even the effective practical performance) and theoretical guarantees. Therefore, we are naturally led to the following question:

In what context can old OCO algorithms work under heavy tails, in what sense, and to what extent?

1.1 Contributions

Motivated by the above question, we examine three classical algorithms for OCO: Online Gradient Descent (OGD) [54], Dual Averaging (DA) [27, 47], and AdaGrad [10, 25], and answer it as follows:

Under the standard bounded domain assumption, the in-expectation regret $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right]$ *is finite and optimal for any* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$*, without any algorithmic modification.*

In detail, our new results for heavy-tailed OCO are summarized here:

- We prove the only and the first optimal regret bound $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}$ for any $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$. Remarkably, AdaGrad can achieve this result without knowing any of the Lipschitz parameter G, noise level σ, and tail index p.

- We extend the analysis of OGD to Online Strongly Convex Optimization with heavy tails and establish the first provable result $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \frac{G^2 \log T}{\mu} + \frac{\sigma^{\mathsf{p}} G^{2-\mathsf{p}}}{\mu} T^{2-\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}$, where $\mu > 0$ is the modulus of strong convexity and T^0 should be read as $\log T$.

Based on the new regret bounds for OCO with heavy tails, we provide the following applications:

- For nonsmooth convex optimization with heavy tails, we show the first optimal in-expectation rate $GD/\sqrt{T} + \sigma D / T^{1-1/\mathsf{p}}$ achieved without gradient clipping, which applies to both the average iterate and last iterate, demonstrating that SGD does converge once the domain is bounded.

- For nonsmooth nonconvex optimization with heavy tails, we show the first provable sample complexity of $G^2 \delta^{-1} \epsilon^{-3} + \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}} \delta^{-1} \epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ for finding a (δ, ϵ)-stationary point without gradient clipping. In addition, we also establish the first lower bound for nonsmooth nonconvex optimization under heavy tails, matching our sample complexity in the high accuracy and noisy regime (i.e., ϵ is small enough with $\sigma > 0$). These two results together provide a nearly complete characterization of the complexity of finding (δ, ϵ)-stationary points in the heavy-tailed setting. Moreover, we give the first convergence result when the problem-dependent parameters (like G, σ, and p) are unknown in advance, resolving a question asked by [20].

Furthermore, we explore broader settings. For example, when each ℓ_t is H-smooth, we present new regrets that extend the classical $L^\star$ bounds [29] to the heavy-tailed noise case. As an important implication, we show that SGD converges at a rate of $HD^2/T + \sigma D / T^{1-1/\mathsf{p}}$ for smooth convex optimization with heavy tails. Finally, we extend our ideas to optimistic algorithms to address various cases simultaneously and employ optimistic algorithms to give the first provable result for Hölder smooth nonconvex optimization under heavy tails, where the problem-dependent parameters can be either known or unknown.

1.2 Discussion on [51]

As noted, [51] is the only work for OCO with heavy tails, as far as we know. There are two major discrepancies between them and us. First, they consider the case where the feasible set $\mathcal{X}$ is unbounded and aim to establish a parameter-free regret bound, i.e., the regret bound has a linear dependency on $\|\boldsymbol{x}\|$ (up to an extra polylog $\|\boldsymbol{x}\|$) for any competitor $\boldsymbol{x} \in \mathcal{X}$. Second, they focus on high-probability rather than in-expectation analysis. As such, their regret is in the form of $\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}) \lesssim (G+\sigma)\|\boldsymbol{x}\| T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}$ (up to extra polylogarithmic factors) with high probability. Without a doubt, their setting is harder than ours implying their bound is stronger as it can convert to an in-expectation regret $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim (G+\sigma)DT^{1/\mathsf{p}}$ for any bounded domain $\mathcal{X}$ with a diameter D.

We emphasize that the motivation behind [51] differs heavily from ours. They aim to solve heavy-tailed OCO with a new proposed method that contains many nontrivial technical tricks, including gradient clipping, artificially added regularization, and solving the additional fixed-point equation. However, their result cannot reflect why the existing simple OCO algorithms like OGD work in practice under heavy-tailed noise. In contrast, our goal is to examine whether, when, and how the classical OCO algorithms work under heavy tails, thereby filling the missing piece in the literature.

Moreover, we would like to mention two drawbacks of [51]. First, though the $T^{1/\mathsf{p}}$ regret seems tight as it matches the lower bound [26, 32, 45], this may not be the best, since an optimal bound should recover the standard $\sqrt{T}$ regret in the deterministic case (i.e., $\sigma = 0$), as one can imagine. This suggests that their bound is not entirely optimal. Second, we remark that they require knowing both problem-dependent parameters G, σ, p and time horizon T in the algorithm, which may be hard to satisfy in the online setting. In comparison, our regret bound $GD\sqrt{T} + \sigma DT^{1/\mathsf{p}}$ is fully optimal in all parameters. Importantly, AdaGrad can achieve it while oblivious to the problem information.

2 Preliminary

Notation. $\mathbb{N}$ denotes the set of natural numbers (excluding 0). $[T] \triangleq \{1,\dots,T\}, \forall T \in \mathbb{N}$. $a \wedge b \triangleq \min\{a,b\}$ and $a \vee b \triangleq \max\{a,b\}$. We write $a \lesssim b$ (resp., $a \gtrsim b$) if $a \le Cb$ (resp., $a \ge Cb$) for a universal constant $C > 0$. $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ respectively represent the floor and ceiling functions. $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product and $\|\cdot\| \triangleq \sqrt{\langle\cdot,\cdot\rangle}$ is the standard 2-norm. Given $\boldsymbol{x} \in \mathbb{R}^d$ and $D > 0$, $\mathcal{B}^d(\boldsymbol{x}, D)$ is the Euclidean ball in $\mathbb{R}^d$ centered at $\boldsymbol{x}$ with a radius D. In the case $\boldsymbol{x} = \mathbf{0}$, we use the shorthand $\mathcal{B}^d(D)$. Given $A \subseteq \mathbb{R}^d$, $\mathsf{int}A$ stands for the interior points of A. For nonempty closed convex $A \subseteq \mathbb{R}^d$, Π_A is the Euclidean projection operator onto A. For a convex function f, $\partial f(\boldsymbol{x})$ denotes its subgradient set at $\boldsymbol{x}$.

Remark 1. We choose the Euclidean norm only for simplicity. Extending the results in this work to any general norm via Bregman divergence is straightforward.

This work studies OCO in the context of Assumption 1.

Assumption 1. *We consider the following series of assumptions:*

- $\mathcal{X} \subset \mathbb{R}^d$ *is a nonempty closed convex set bounded by D, i.e.,* $\sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{y}\| \le D$.
- $\ell_t : \mathcal{X} \to \mathbb{R}$ *is closed convex for all* $t \in [T]$.
- ℓ_t *is G-Lipschitz on $\mathcal{X}$, i.e.,* $\|\nabla \ell_t(\boldsymbol{x})\| \le G, \forall \boldsymbol{x} \in \mathcal{X}, \nabla \ell_t(\boldsymbol{x}) \in \partial \ell_t(\boldsymbol{x})$, *for all* $t \in [T]$.
- *Given a point $\boldsymbol{x}_t \in \mathcal{X}$ at the t-th iteration, one can query $\boldsymbol{g}_t \in \mathbb{R}^d$ satisfying* $\nabla \ell_t(\boldsymbol{x}_t) \triangleq \mathbb{E}[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}] \in \partial \ell_t(\boldsymbol{x}_t)$ *and* $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$ *for some* $\mathsf{p} \in (1,2]$ *and* $\sigma \ge 0$, *where* $\mathcal{F}_t \triangleq \sigma(\boldsymbol{g}_1,\dots,\boldsymbol{g}_t)$ *denotes the natural filtration and* $\boldsymbol{\epsilon}_t \triangleq \boldsymbol{g}_t - \nabla \ell_t(\boldsymbol{x}_t)$ *is the stochastic noise.*

Remark 2. D is recognized as known, like ubiquitously assumed in the OCO literature. Moreover, $\boldsymbol{x}_t$ denotes the decision/output of the online learning algorithm by default.

In Assumption 1, the first three points are standard, and the fourth is the heavy-tailed noise assumption. In particular, $\mathsf{p} = 2$ recovers the standard finite variance condition.

3 Old Algorithms under Heavy Tails

In this section, we revisit three classical algorithms for OCO: OGD, DA, and $\mathsf{AdaGrad}$, whose regret bounds are well-studied in the finite variance case but remain unknown under heavy-tailed noise.

The basic idea of proving these algorithms work under heavy tails is to leverage the boundness property of $\mathcal{X}$. We will describe it in more detail using OGD as an illustrated example. The analysis of DA follows a similar way at a high level, but differs in some details. However, though $\mathsf{AdaGrad}$ can be viewed as OGD with an adaptive stepsize, the way to utilize the boundness property is entirely different. All formal proofs are deferred to the appendix due to space limitations.

3.1 New Regret for Online Gradient Descent

Algorithm 1 Online Gradient Descent (OGD) [54]

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, stepsize $\eta_t > 0$
for $t = 1$ **to** T **do**
  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t)$
end for

We begin from arguably the most basic algorithm for OCO, Online Gradient Descent (OGD).

A well known analysis. The regret bound of OGD has been extensively studied [11, 29, 38]. The most well known analysis is perhaps the following one: for any $\boldsymbol{x} \in \mathcal{X}$, there is

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 = \|\Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) - \Pi_{\mathcal{X}}(\boldsymbol{x})\|^2 \le \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{x}\|^2,$$

where the inequality holds by the nonexpansive property of $\Pi_{\mathcal{X}}$. Expanding both sides and rearranging terms yield that

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \frac{\eta_t \|\boldsymbol{g}_t\|^2}{2}. \tag{1}$$

If $\boldsymbol{g}_t$ admits a finite variance, i.e., $\mathsf{p} = 2$ in Assumption 1, taking expectations on both sides, then following a standard analysis for $\eta_t = \frac{D}{(G+\sigma)\sqrt{t}}$ (or $\eta_t = \frac{D}{(G+\sigma)\sqrt{T}}$ if T is known) gives the regret

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim (G+\sigma) D\sqrt{T}, \forall \boldsymbol{x} \in \mathcal{X}.$$

However, the step of taking expectations on the R.H.S. of (1) crucially relies on the finite variance condition of $\boldsymbol{g}_t$. Therefore, one may naturally think OGD would not guarantee a finite regret if $\mathsf{p} < 2$.

A less well known analysis[1]**.** As discussed, the failure of the above proof under heavy-tailed noise is due to (1). Therefore, if a tighter inequality than (1) exists, then it might be possible to show that OGD still works for $\mathsf{p} < 2$. However, does it exist?

Actually, there is another less well known analysis to produce a better inequality than (1). That is, first showing for any $\boldsymbol{x} \in \mathcal{X}$, by the optimality condition of the update rule,

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle \le \frac{\langle \boldsymbol{x}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle}{\eta_t} = \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t},$$

and then obtaining

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t}. \tag{2}$$

Note that (2) is tighter than (1) as $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \le \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le \frac{\eta_t \|\boldsymbol{g}_t\|^2}{2} + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t}$, where the first step is due to Cauchy-Schwarz inequality and the second one is by AM-GM inequality.

[1] To clarify, the phrase "less well known" is compared to the first one. This analysis itself is also well known.

Handle $\mathsf{p} < 2$ in a simple way. Though we have tightened (1) into (2), can inequality (2) help to overcome heavy tails? The answer is surprisingly positive, and our solution is fairly simple. Instead of directly applying AM-GM inequality in the second step, we recall $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{x}_t) + \boldsymbol{\epsilon}_t$ and use triangle inequality to obtain

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \le \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le (\|\nabla \ell_t(\boldsymbol{x}_t)\| + \|\boldsymbol{\epsilon}_t\|) \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|. \tag{3}$$

On the one hand, by $\|\nabla \ell_t(\boldsymbol{x}_t)\| \le G$ and AM-GM inequality, there is

$$\|\nabla \ell_t(\boldsymbol{x}_t)\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le G \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le \eta_t G^2 + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t}. \tag{4}$$

On the other hand, let $\mathsf{p}_\star \triangleq \frac{\mathsf{p}}{\mathsf{p}-1}$ and $\mathsf{C}(\mathsf{p}) \triangleq \frac{(4\mathsf{p}-4)^{\mathsf{p}-1}}{\mathsf{p}^\mathsf{p}}$, we have

$$\begin{aligned}
\|\boldsymbol{\epsilon}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| &= \left(\frac{4\eta_t}{\mathsf{p}_\star}\right)^{\frac{1}{\mathsf{p}_\star}} \|\boldsymbol{\epsilon}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^{1-\frac{2}{\mathsf{p}_\star}} \cdot \left(\frac{\mathsf{p}_\star \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t}\right)^{\frac{1}{\mathsf{p}_\star}} \\
&\overset{(a)}{\le} \frac{\left(\frac{4\eta_t}{\mathsf{p}_\star}\right)^{\frac{\mathsf{p}}{\mathsf{p}_\star}} \|\boldsymbol{\epsilon}_t\|^\mathsf{p} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^{\mathsf{p}-\frac{2\mathsf{p}}{\mathsf{p}_\star}}}{\mathsf{p}} + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t} \\
&\overset{(b)}{\le} \mathsf{C}(\mathsf{p}) \eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^\mathsf{p} D^{2-\mathsf{p}} + \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{4\eta_t},
\end{aligned} \tag{5}$$

where (a) is by Young's inequality and (b) is due to $\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le D$, $\mathsf{p}_\star = \frac{\mathsf{p}}{\mathsf{p}-1}$, and $\mathsf{C}(\mathsf{p}) = \frac{(4\mathsf{p}-4)^{\mathsf{p}-1}}{\mathsf{p}^\mathsf{p}}$. Next, we plug (4) and (5) back into (3), then combine with (2) to know

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p}) \eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^\mathsf{p} D^{2-\mathsf{p}}. \tag{6}$$

Notably, the term $\|\boldsymbol{\epsilon}_t\|^\mathsf{p}$ has a correct exponent p. Thus, we can safely take expectations on both sides. Finally, a standard analysis yields the following Theorem 1 (see Appendix A for a formal proof).

Theorem 1. *Under Assumption 1, taking* $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ *in* OGD *(Algorithm 1), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

As far as we know, Theorem 1 is the first and the only provable result for OGD under heavy tails. Remarkably, it is not only tight in T [26, 32, 45] but also fully optimal in all parameters, in contrast to the bound $(G+\sigma)DT^{1/\mathsf{p}}$ of [51]. This reveals that OCO with heavy tails can be optimally solved as effectively as the finite variance case once the domain is bounded, a classical condition adapted in many existing works.

Strongly convex functions. We highlight that the above idea can also be applied to Online Strongly Convex Optimization and leads to a sublinear regret $T^{2-\mathsf{p}}$ better than $T^{1/\mathsf{p}}$. This extension can be found in Appendix A.

3.2 New Regret for Dual Averaging

Algorithm 2 Dual Averaging (DA) [27, 47]

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, stepsize $\eta_t > 0$
for $t = 1$ **to** T **do**
  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_1 - \eta_t \sum_{s=1}^t \boldsymbol{g}_s)$
end for

Remark 3. It is known that DA is a special realization of the more general Follow-the-Regularized-Leader (FTRL) framework [24]. To keep the work concise, we only focus on DA. The key idea to prove Theorem 2 can directly extend to show new regret for FTRL under heavy-tailed noise.

We turn our attention to the second candidate, the Dual Averaging (DA) algorithm, which is given in Algorithm 2. Though DA coincides with OGD when $\mathcal{X} = \mathbb{R}^d$ and $\eta_t = \eta$, these two methods in general are not equivalent and can have significant performance differences in practice. Therefore, it is also important to understand DA under heavy tails.

Despite the proof strategies for OGD and DA are in different flavors (even for $\mathsf{p} = 2$), the basic idea presented before for OGD still works here, i.e., apply the boundness property of $\mathcal{X}$ to make the term $\|\boldsymbol{\epsilon}_t\|$ have a correct exponent. Armed with this thought, we can prove the following new regret bound for DA under heavy-tailed noise. We refer the reader to Appendix B for its proof.

Theorem 2. *Under Assumption 1, taking $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ in* DA *(Algorithm 2), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

As far as we know, Theorem 2 is the first provable and optimal regret for DA under heavy tails. It guarantees the same tight bound as in Theorem 1 up to different constants.

3.3 New Regret for AdaGrad

Algorithm 3 AdaGrad [10, 25]

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, stepsize $\eta > 0$
for $t = 1$ **to** T **do**
  $\eta_t = \eta V_t^{-1/2}$ where $V_t = \sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2$
  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t)$
end for

Remark 4. Algorithm 3 is also named AdaGrad-Norm (e.g., [46]). We simply call it AdaGrad. It is straightforward to generalize Theorem 3 below to the per-coordinate update version.

Although Theorems 1 and 2 are optimal, they both suffer from an undesired point. That is, the stepsize $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ requires knowing all problem-dependent parameters. However, it may not be easy to obtain them in an online setting. Especially, it heavily depends on the prior information about the tail index p, which is hard to know (even approximately) in advance. In other words, they both lack the adaptive property to an unknown environment.

To handle this issue, we consider AdaGrad, a classical adaptive algorithm for OCO. As can be seen, AdaGrad is just OGD with an adaptive stepsize. However, it is this adaptive stepsize that can help us to overcome the above undesired point.

Theorem 3. *Under Assumption 1, taking $\eta = D/\sqrt{2}$ in* AdaGrad *(Algorithm 3), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{AdaGrad}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Remark 5. We also establish a similar result for DA with an adaptive stepsize. See Theorem 9 in Appendix B for details.

Theorem 3 provides the first regret bound for AdaGrad under heavy tails. Impressively, it is optimal even without knowing any of G, σ, and p. This surprising result once again demonstrates the power of the adaptive method, indicating it is robust to an unknown environment and even heavy-tailed noise, which may partially explain the favorable performance of many adaptive optimizers designed based on AdaGrad like RMSProp [44] and Adam [17].

We point out that the key to establishing Theorem 3 differs from the idea used before for OGD and DA. Actually, Theorem 3 can be obtained in an embarrassingly simple way. It is known that AdaGrad with $\eta = D/\sqrt{2}$ on a bounded domain guarantees the following path-wise regret

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \lesssim D\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2}. \tag{7}$$

Observe that $\sqrt{\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2} \lesssim \sqrt{\sum_{t=1}^{T}\|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \sqrt{\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^2} \le G\sqrt{T} + \left(\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}$, where the last step is due to $\|\cdot\|_2 \le \|\cdot\|_{\mathsf{p}}$ for any $\mathsf{p} \in [1,2]$. After taking expectations on both sides of (7) and applying Hölder's inequality to obtain $\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \le \left(\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \le \sigma T^{\frac{1}{\mathsf{p}}}$, we conclude Theorem 3. To make the work self-consistent, we produce the formal proof of Theorem 3 in Appendix C.

4 Applications

We provide some applications based on the new regret bounds established in Section 3. The basic problem we study is optimizing a single objective F, which could be either convex or nonconvex.

4.1 Nonsmooth Convex Optimization

In this section, we consider nonsmooth convex optimization with heavy tails.

Convergence of the average iterate. First, we focus on convergence in average. By the classical online-to-batch conversion [4], the following corollary immediately holds.

Corollary 1. *Under Assumption 1 for* $\ell_t(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}\rangle$ *and let* $\bar{\boldsymbol{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t$*, for any* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$*, we have*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x})\right] \le \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right]}{T} \lesssim \frac{GD}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. By convexity, $F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}) \le \frac{\sum_{t=1}^{T} F(\boldsymbol{x}_t) - F(\boldsymbol{x})}{T} \le \frac{\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})}{T}$ is valid for any OCO algorithm A. We conclude from invoking Theorems 1, 2 and 3. □

To the best of our knowledge, Corollary 1 gives the first and optimal convergence rate for these three algorithms in stochastic optimization with heavy tails. Especially, it implies that once the domain is bounded, the widely implemented SGD algorithm provably converges under heavy-tailed noise without any algorithmic change considered in many prior works, e.g., gradient clipping [21, 28].

We are only aware of two works [22, 45] based on Stochastic Mirror Descent (SMD) [26] that gave convergence results without clipping. However, they share a common drawback, i.e., their bounds are both in the form of $(G+\sigma)D/T^{1-1/\mathsf{p}}$, which cannot recover the optimal rate $GD/\sqrt{T}$ when $\sigma = 0$.

Lastly, we highlight that for $\mathsf{A} = \mathsf{AdaGrad}$, Corollary 1 is not only optimal but also adaptive to the tail index p. As far as we know, no result has achieved this property before. This once again evidences the benefit of adaptive gradient methods.

Convergence of the last iterate. Next, we consider the more challenging last-iterate convergence, which has a long history in stochastic optimization and fruitful results in the case of $\mathsf{p} = 2$ (see, e.g., [30, 39, 53]). However, less is known about heavy-tailed problems. So far, only two works [22, 31] have established the last-iterate convergence. The former is based on SMD, and the latter employs gradient clipping in SGD. Unfortunately, their rates are both in the suboptimal order $(G+\sigma)D/T^{1-1/\mathsf{p}}$.

We will provide an optimal last-iterate rate based on the following lemma, which reduces the last-iterate convergence to an online learning problem.

Lemma 1 (Theorem 1 of [9])**.** *Suppose* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ *and* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ *are two sequences of vectors satisfying* $\boldsymbol{x}_t \in \mathcal{X}$, $\boldsymbol{x}_1 = \boldsymbol{y}_1$ *and*

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \frac{T-t}{T}\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\right). \tag{8}$$

Given a convex function $F(\boldsymbol{x})$*, let* $\ell_t(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{y}_t), \boldsymbol{x}\rangle$*. Then for any online learner* A*, we have*

$$F(\boldsymbol{y}_T) - F(\boldsymbol{x}) \le \frac{\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})}{T}, \forall \boldsymbol{x} \in \mathcal{X}.$$

We emphasize that the stochastic gradient $\boldsymbol{g}_t$ received by A is an estimate of $\nabla F(\boldsymbol{y}_t)$ instead of $\nabla F(\boldsymbol{x}_t)$. This flexibility is due to the generality of the OCO framework. Moreover, for OGD, suppose there is no projection step, then (8) is equivalent to $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t - \frac{T-t}{T}\eta_t \boldsymbol{g}_t$, which can be viewed as SGD with a stepsize $\frac{T-t}{T}\eta_t$. For proof of Lemma 1, we refer the interested reader to [9].

Corollary 2. *Under Assumption 1 for $\ell_t(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{y}_t), \boldsymbol{x}\rangle$, where $\boldsymbol{y}_t$ satisfies (8), for any $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$, we have*

$$\mathbb{E}\left[F(\boldsymbol{y}_T) - F(\boldsymbol{x})\right] \le \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right]}{T} \lesssim \frac{GD}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. Combine Lemma 1 and Theorems 1, 2 and 3 to conclude. □

As far as we know, Corollary 2 is the first optimal last-iterate convergence rate for stochastic convex optimization with heavy tails, closing the gap in existing works.

One may notice that $\boldsymbol{y}_t$ itself is not the decision made by the online learner and naturally may ask whether $\boldsymbol{x}_t$ ensures the last-iterate convergence if we simply pick $\ell_t = F$. The answer turns out to be positive at least for OGD (which is equivalent to SGD now). However, to prove this result, we rely on a technique specialized to stochastic optimization recently developed by [22, 48]. To not diverge from the topic of OCO, we defer the last-iterate convergence of OGD to Appendix D, in which Theorem 10 gives a general result for any stepsize η_t and Corollary 5 shows a last-iterate rate similar to Corollary 2 (up to an extra logarithmic factor) under the same stepsize $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ as in Theorem 1.

4.2 Nonsmooth Nonconvex Optimization

This section contains another application, nonsmooth nonconvex optimization with heavy tails. Due to limited space, we will provide only the necessary background. For more details, we refer the reader to [8, 15, 18, 19, 42, 43] for recent progress. We start with a new set of conditions.

Assumption 2. *We consider the following series of assumptions:*

- *The objective F is lower bounded by $F_\star \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) \in \mathbb{R}$.*
- *F is differentiable and well-behaved, i.e., $F(\boldsymbol{x}) - F(\boldsymbol{y}) = \int_0^1 \langle \nabla F(\boldsymbol{y} + t(\boldsymbol{x}-\boldsymbol{y})), \boldsymbol{x}-\boldsymbol{y}\rangle \,\mathrm{d}t$.*
- *F is G-Lipschitz on $\mathbb{R}^d$, i.e., $\|\nabla F(\boldsymbol{x})\| \le G, \forall \boldsymbol{x} \in \mathbb{R}^d$.*
- *Given $\boldsymbol{z}_t \in \mathbb{R}^d$ at the t-th iteration, one can query $\boldsymbol{g}_t \in \mathbb{R}^d$ satisfying $\mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right] = \nabla F(\boldsymbol{z}_t)$ and $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$ for some $\mathsf{p} \in (1,2]$ and $\sigma \ge 0$, where $\mathcal{F}_t$ denotes the natural filtration and $\boldsymbol{\epsilon}_t \triangleq \boldsymbol{g}_t - \nabla F(\boldsymbol{z}_t)$ is the stochastic noise.*

Remark 6. The second point is a mild regularity condition introduced by [7] and becomes standard in the literature [2, 20, 52]. See Definition 1 and Proposition 2 of [7] for more details. In the fourth point, we use the same notation $\boldsymbol{z}_t$ as in the algorithm being studied later. In fact, it can be arbitrary.

In nonsmooth nonconvex optimization, we aim to find a (δ, ϵ)-stationary point [50] (see the formal Definition 2 in Appendix E). This goal can be reduced to finding a point $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\nabla F(\boldsymbol{x})\|_\delta \le \epsilon$, where $\|\nabla F(\boldsymbol{x})\|_\delta$ is a quantity introduced by [7] as follows.

Definition 1 (Definition 5 of [7])**.** Given a point $\boldsymbol{x} \in \mathbb{R}^d$, a number $\delta > 0$ and an almost-everywhere differentiable function F, define $\|\nabla F(\boldsymbol{x})\|_\delta \triangleq \inf_{S\subset\mathcal{B}(\boldsymbol{x},\delta), \frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\boldsymbol{y}=\boldsymbol{x}} \left\| \frac{1}{|S|}\sum_{\boldsymbol{y}\in S} \nabla F(\boldsymbol{y})\right\|$.

The only existing sample complexity under Assumption 2 is $(G+\sigma)^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ in high probability [20], where we only report the dominant term and hide the dependency on the failure probability.

However, on the theoretical side, their result cannot recover the optimal bound $G^2\delta^{-1}\epsilon^{-3}$ [7] in the deterministic case. On the practical side, their method also employs the gradient clipping step, which introduces a new clipping parameter to tune. In fact, as stated in their Section 5, they observed in experiments that their algorithm without the clipping operation (exactly the algorithm we study next) still works under heavy tails. In addition, in their Section 6, they also explicitly ask whether the requirement to know G and A can be removed.

As will be seen later, we can address these points with the new regret bounds presented before.

4.2.1 Online-to-Nonconvex Conversion under Heavy Tails

Algorithm 4 Online-to-Nonconvex Conversion (O2NC) [7]

Input: initial point $\boldsymbol{y}_0 \in \mathbb{R}^d$, $K \in \mathbb{N}$, $T \in \mathbb{N}$, online learning algorithm A.
for $n = 1$ **to** KT **do**
  Receive $\boldsymbol{x}_n$ from A
  $\boldsymbol{y}_n = \boldsymbol{y}_{n-1} + \boldsymbol{x}_n$
  $\boldsymbol{z}_n = \boldsymbol{y}_{n-1} + s_n \boldsymbol{x}_n$ where $s_n \sim \mathsf{Uniform}\,[0,1]$ i.i.d.
  Query a stochastic gradient $\boldsymbol{g}_n$ at $\boldsymbol{z}_n$
  Send $\boldsymbol{g}_n$ to A
end for

Remark 7. Note that O2NC is a randomized algorithm. Therefore, the definition of the natural filtration is adjusted to $\mathcal{F}_n \triangleq \sigma(s_1, \boldsymbol{g}_1, \ldots, s_n, \boldsymbol{g}_n, s_{n+1})$ accordingly.

We provide the Online-to-Nonconvex Conversion (O2NC) framework in Algorithm 4, which serves as a meta algorithm. Roughly speaking, Algorithm 4 reduces a nonconvex optimization problem to an OCO (in fact, OLO) problem, for which the K-shifting regret (see (9)) of the online learner A crucially affects the final convergence rate. However, the existing Theorem 8 of [7], a general convergence result for the above reduction, cannot directly apply to heavy-tailed noise, since its proof relies on the finite variance condition on $\boldsymbol{g}_n$ (see Appendix E for more details).

Theorem 4. *Under Assumption 2 and let* $\boldsymbol{v}_k \triangleq -D\frac{\sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n)}{\left\|\sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n)\right\|}, \forall k \in [K]$ *for arbitrary* $D > 0$, *then for any online learning algorithm* A *in* O2NC *(Algorithm 4), we have*

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{K}\left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n) \right\|\right] \lesssim \frac{F(\boldsymbol{y}_0) - F_\star}{DKT} + \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)\right]}{DKT} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$ in Theorem 4 is called *K-shifting regret* [7], defined as follows:

$$\mathsf{R}_T^{\mathsf{A}}\left(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\right) \triangleq \sum_{k=1}^{K} \sum_{n=(k-1)T+1}^{kT} \ell_n(\boldsymbol{x}_n) - \ell_n(\boldsymbol{v}_k) \quad \text{where} \quad \ell_n(\boldsymbol{x}) \triangleq \langle \boldsymbol{g}_n, \boldsymbol{x} \rangle . \tag{9}$$

Theorem 4 here provides a new and the first theoretical guarantee for O2NC under heavy tails. Especially, it recovers Theorem 8 of [7] when $\mathsf{p} = 2$. A remarkable point is that the O2NC algorithm itself does not need any information about p. The proof of Theorem 4 can be found in Appendix E.

4.2.2 Convergence Rates

Theorem 4 enables us to apply the results presented in Section 3. Concretely, for $\mathcal{X} = \mathcal{B}^d(D)$ and any $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$, if we reset the stepsize in A after every T iterations, there will be $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)\right] \lesssim GDK\sqrt{T} + \sigma DKT^{1/\mathsf{p}}$ by our new regret bounds, since $\boldsymbol{v}_k \in \mathcal{X}$. With a carefully picked D, we obtain the following Theorem 5. Its proof is deferred to Appendix E.

Theorem 5. *Under Assumption 2 and let* $\Delta \triangleq F(\boldsymbol{y}_0) - F_\star$ *and* $\bar{\boldsymbol{z}}_k \triangleq \frac{1}{T}\sum_{n=(k-1)T+1}^{kT} \boldsymbol{z}_n, \forall k \in [K]$, *setting any* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$ *in* O2NC *(Algorithm 4) with a domain* $\mathcal{X} = \mathcal{B}^d(D)$ *for* $D = \delta/T$ *and resetting the stepsize in* A *after every* T *iterations, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\Delta}{\delta K} + \frac{G}{\sqrt{T}} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

Notably, this is the first time confirming that gradient clipping is indeed unnecessary for the O2NC framework, matching the experimental observation of [20].

Corollary 3. *Under the same setting of Theorem 5, suppose we have $N \geq 2$ stochastic gradient budgets, taking $K = \lfloor N/T \rfloor$ and $T = \lceil N/2 \rceil \wedge \left(\left\lceil (\delta G N/\Delta)^{\frac{2}{3}} \right\rceil \vee \left\lceil (\delta \sigma N/\Delta)^{\frac{\mathsf{p}}{2\mathsf{p}-1}} \right\rceil \right)$, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{G}{\sqrt{N}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \frac{\Delta}{\delta N} + \frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}} + \frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

Corollary 3 is obtained by optimizing K and T in Theorem 5. It implies a sample complexity of $G^2\delta^{-1}\epsilon^{-3} + \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ for finding a (δ,ϵ)-stationary point, improved over the previous bound $(G+\sigma)^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ [20]. Furthermore, leveraging the adaptive feature of AdaGrad, Corollary 6 in Appendix E shows how to set K and T without G, σ, and p, resulting in the first provably rate for O2NC when no problem information is known in advance, which solves the problem asked by [20].

4.2.3 Lower Bounds

In this part, we provide the first lower bound for nonsmooth nonconvex optimization under heavy tails in the following Theorem 6, the proof of which follows the framework first established in [3] and later developed by [7] but with some necessary (though minor) variation to make it compatible with heavy-tailed noise.

Theorem 6. *For any given $\Delta > 0$, $G > 0$, $\mathsf{p} \in (1,2]$, $\sigma \geq 0$, $\delta > 0$ and $0 < \epsilon \lesssim \frac{\Delta}{\delta} \wedge \frac{G^2\delta}{\Delta}$, there exists a dimension $d > 0$ depending on the previous parameters such that, for any randomized first-order algorithm (see Definition 4 for a formal description), there exists a G-Lipschitz differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ satisfying $F(\mathbf{0}) - F_\star \leq \Delta$ and a function $\boldsymbol{g} : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}$ satisfying $\mathbb{E}_r[\boldsymbol{g}(\boldsymbol{x},r)] = \nabla F(\boldsymbol{x})$ and $\mathbb{E}_r\left[\|\boldsymbol{g}(\boldsymbol{x},r) - \nabla F(\boldsymbol{x})\|^{\mathsf{p}}\right] \leq \sigma^{\mathsf{p}}$ where r follows a certain probability distribution over $\{0,1\}$ such that the algorithm requires $\gtrsim \Delta\delta^{-1}\epsilon^{-1} + \Delta\sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$ queries of $\boldsymbol{g}$ to find a point $\boldsymbol{z}$ such that $\mathbb{E}[\|\nabla F(\boldsymbol{z})\|_\delta] \leq \epsilon$.*

For small enough ϵ and $\sigma > 0$, Theorem 6 can be further simplified into a lower bound of $\Delta\sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$, matching the leading term in the sample complexity derived from the previous Corollary 3. Therefore, Theorem 6 suggests that our Corollary 3 is tight in the high accuracy and noisy regime. As such, Corollary 3 and Theorem 6 together provide a nearly complete characterization of the complexity of finding (δ,ϵ)-stationary points in the heavy-tailed setting.

However, for any general $\epsilon > 0$ or the case $\sigma = 0$, there is still a gap between the upper and lower bounds. Closing this gap could be an interesting direction for the future.

5 Further Extensions

So far, we revisit three classical OCO algorithms, OGD, DA, and AdaGrad, show that they all guarantee optimal regrets, and provide some applications based on their new bounds. However, all the results are restricted to the Lipschitz case, which is standard in the literature, but sometimes inadequate to derive a better bound. For example, if ℓ_t is smooth (i.e., the gradient of ℓ_t is Lipschitz), then one could establish a bound depending on the cumulative competitor loss, i.e., $\sum_{t=1}^{T}\ell_t(\boldsymbol{x})$.

In this section, we first show that our idea presented in Section 3 can be directly extended to the smooth case, then discuss what more we can do. Formally, we need the following condition.

Condition 1. ℓ_t is H-smooth on $\mathbb{R}^d$, i.e., $\|\nabla\ell_t(\boldsymbol{x}) - \nabla\ell_t(\boldsymbol{y})\| \leq H\|\boldsymbol{x}-\boldsymbol{y}\|, \forall \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d$, for all $t \in [T]$.

Remark 8. Strictly speaking, Condition 1 may not be well-defined, since Assumption 1 only requires ℓ_t to be defined on $\mathcal{X}$. To avoid any further complications, this section recognizes ℓ_t as a real-valued convex function defined on $\mathbb{R}^d$, and $\mathcal{X}$ is the constraint set, i.e., the domain of the problem.

5.1 New Regrets for Old Algorithms under Smooth ℓ_t and Applications

Theorem 7. *Under Assumption 1 (with replacing the third point by Condition 1) and additionally assuming $\ell_t \geq 0$ on $\mathbb{R}^d$:*

- *taking* $\eta_t = \frac{1}{4H} \wedge \frac{\gamma D}{\sqrt{H}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ *for any* $\gamma > 0$ *in* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$*, we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim HD^2 + \sqrt{H}D\left(\frac{1}{\gamma} + \gamma\sum_{t=1}^{T}\ell_t(\boldsymbol{x})\right) + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

- *taking* $\eta = D/\sqrt{2}$ *in* $\mathsf{A} = \mathsf{AdaGrad}$*, we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim HD^2 + D\sqrt{H\sum_{t=1}^{T}\ell_t(\boldsymbol{x})} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Theorem 7 provides new regrets that extend the classical $L^\star$ bounds [29] to the heavy-tailed noise case, under an additional nonnegative condition $\ell_t \geq 0$ that is widely used in the literature [41]. The full version without such a requirement is provided in Theorem 13 in Appendix F.

As one can see, the optimal value of γ for OGD and DA should be in the order of $1/\sqrt{\sum_{t=1}^{T}\ell_t(\boldsymbol{x})}$, which is however not possible to take as the competitor $\boldsymbol{x}$ is not fixed here. In contrast, $\mathsf{AdaGrad}$ contains the improved term $D\sqrt{H\sum_{t=1}^{T}\ell_t(\boldsymbol{x})}$, which once again suggests the benefit of adaptive methods. More importantly, the regret for $\mathsf{AdaGrad}$ also indicates that it can be oblivious to the problem class and adapt to the best-possible bound automatically, even under heavy tails.

Theorem 7 could be particularly useful for smooth convex optimization when a global minimizer is contained in $\mathcal{X}$, as given in the following corollary.

Corollary 4. *Under Assumption 1 (with replacing the third point by Condition 1) for* $\ell_t(\boldsymbol{x}) = F(\boldsymbol{x})$ *and additionally assuming that* $\mathcal{X}$ *contains a global minimzer* $\boldsymbol{x}^\star$ *of* F *(i.e.,* $\boldsymbol{x}^\star \in \operatorname{arginf}_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x})$*), let* $\bar{\boldsymbol{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t$*, taking* $\eta_t = \frac{1}{4H} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ *in* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$ *or* $\eta = D/\sqrt{2}$ *in* $\mathsf{A} = \mathsf{AdaGrad}$*, we have*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}^\star)\right] \leq \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}^\star)\right]}{T} \lesssim \frac{HD^2}{T} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}.$$

Proof. Equivalently, we can redefine $\ell_t(\boldsymbol{x}) = F(\boldsymbol{x}) - F(\boldsymbol{x}^\star)$, then ℓ_t is nonnegative on $\mathbb{R}^d$. By convexity, $F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}^\star) \leq \frac{\sum_{t=1}^{T} F(\boldsymbol{x}_t) - F(\boldsymbol{x}^\star)}{T} = \frac{\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}^\star)}{T}$ is valid for any OCO algorithm A. We conclude from invoking Theorem 7 with $\gamma = \frac{1}{4D\sqrt{H}}$ for $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$ and $\eta = D/\sqrt{2}$ for $\mathsf{A} = \mathsf{AdaGrad}$ and combining the fact $\ell_t(\boldsymbol{x}^\star) = 0$. □

An important implication of Corollary 4 is the first to show that SGD can converge at the optimal rate of $HD^2/T + \sigma D/T^{1-1/\mathsf{p}}$ for heavy-tailed smooth convex optimization (though under an extra condition). In fact, a valid but worse bound without assuming the existence of $\boldsymbol{x}^\star \in \mathcal{X}$ is possible, which we defer to Corollary 7 in Appendix F.

5.2 Optimistic Algorithms for Broader Cases

As presented, we have successfully handled smooth OCO with heavy tails for classical OGD, DA, and $\mathsf{AdaGrad}$. In fact, our ideas can also be applied to another famous family of methods known as optimistic algorithms [6, 16, 33] to deal with broader cases. Specifically, we will study an optimistic version of $\mathsf{AdaGrad}$ called $\mathsf{OAdaGrad}$ in Appendix F to address various cases simultaneously (see Theorems 14 and 15) and provide more applications for stochastic optimization in Appendix G.

For example, in the special case of heavy-tailed Hölder smooth optimization (i.e., $\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq H\|\boldsymbol{x} - \boldsymbol{y}\|^\nu$ for some $H > 0$ and $\nu \in (0, 1]$), Corollary 8 provides a rate $\frac{HD^{1+\nu}}{T^{\frac{1+\nu}{2}}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}$ for convex problems, Corollaries 11 and 12 establish provable results for nonconvex problems, where the problem-dependent parameters may be known or unknown. All of these results are new to the best of our knowledge.

6 Conclusion and Limitation

This paper shows that three classical OCO algorithms, OGD, DA, and AdaGrad, can achieve the optimal in-expectation regret under heavy tails without any algorithmic modification if the feasible set is bounded, and provides some applications in stochastic optimization. The main limitation of our work is that all the proof crucially relies on the bounded domain assumption, which may not always be suitable in practice. Finding a weaker sufficient condition, under which the classical OCO algorithms work with heavy tails provably, is a direction worth studying in the future.

References

[1] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 111–119. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/agarwal19c.html`.

[2] Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94909–94933. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/ac8ec9b4d94c03f0af8c4fe3d5fad4fd-Paper-Conference.pdf`.

[3] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

[4] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004. doi: 10.1109/TIT.2004.833339.

[5] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[6] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 6.1–6.20, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL `https://proceedings.mlr.press/v23/chiang12.html`.

[7] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6643–6670. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/cutkosky23a.html`.

[8] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6692–6703. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/2c8d9636f74d0207ff4f65956010f450-Paper-Conference.pdf`.

[9] Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

[11] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL http://dx.doi.org/10.1561/2400000013.

[12] Elad Hazan and Karan Singh. Introduction to online control, 2025. URL https://arxiv.org/abs/2211.09619.

[13] Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4262–4274. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hodgkinson21a.html.

[14] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2413–2421. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/hubler25a.html.

[15] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4570–4597. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/jordan23a.html.

[16] Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In Steve Hanneke and Lev Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 681–720. PMLR, 15–17 Oct 2017. URL https://proceedings.mlr.press/v76/joulani17a.html.

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

[18] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022. URL http://jmlr.org/papers/v23/21-1507.html.

[19] Guy Kornowski and Ohad Shamir. On the complexity of finding small subgradients in nonsmooth optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. URL https://openreview.net/forum?id=SaRQ4oTqWbP.

[20] Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32122–32138. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bo.html.

[21] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

[22] Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

[23] Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NKotdPUc3L.

[24] Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 525–533, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `https://proceedings.mlr.press/v15/mcmahan11b.html`.

[25] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, pages 244–256. Omnipress, 2010.

[26] Arkadi Nemirovski and David Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

[27] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120 (1):221–259, 2009.

[28] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/4c454d34f3a4c8d6b4ca85a918e5d7ba-Paper-Conference.pdf`.

[29] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[30] Francesco Orabona. Last iterate of sgd converges (even in unbounded domains). 2020. URL `https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/`.

[31] Daniela Angela Parletta, Andrea Paudice, and Saverio Salzo. An improved analysis of the clipped stochastic subgradient method under heavy-tailed noise, 2025. URL `https://arxiv.org/abs/2410.00573`.

[32] Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 803–510, 2009. doi: 10.1109/ALLERTON.2009.5394945.

[33] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 993–1019, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v30/Rakhlin13.html`.

[34] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[35] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

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

[37] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/sadiev23a.html`.

[38] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012. ISSN 1935-8237. doi: 10.1561/2200000018. URL `http://dx.doi.org/10.1561/2200000018`.

[39] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/shamir13.html.

[40] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/simsekli19a.html.

[41] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

[42] Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.

[43] Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21360–21379. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tian22a.html.

[44] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[45] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 65–102. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/vural22a.html.

[46] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ward19a.html.

[47] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[48] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023.

[49] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

[50] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of*

Machine Learning Research, pages 11173–11182. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhang20p.html.

[51] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8000–8012. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/349956dee974cfdcbbb2d06afad5dd4a-Paper-Conference.pdf.

[52] Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58780–58799. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24k.html.

[53] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[54] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Missing Proofs for Online Gradient Descent

This section provides missing proofs for regret bounds of OGD. Before showing the formal proof, we recall the following core inequality that holds for any $\boldsymbol{x} \in \mathcal{X}$ given in (6):

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \tag{10}$$

The key to establishing the above result is showing

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_t} \le \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}, \tag{11}$$

the proof of which is by combining (3), (4), and (5) established in the main text.

A.1 Proof of Theorem 1

Proof. For any $\boldsymbol{x} \in \mathcal{X}$, sum up (10) from $t = 1$ to T and drop the term $-\frac{\|\boldsymbol{x}_{T+1} - \boldsymbol{x}\|^2}{2\eta_T}$ to obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \\
\le& \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}\|^2}{2\eta_1} + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}} && (12) \\
\le& \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}, && (13)
\end{aligned}$$

where the last step is due to $\|\boldsymbol{x}_t - \boldsymbol{x}\| \le D, \forall t \in [T]$ and $\eta_{t+1} \le \eta_t, \forall t \in [T-1]$.

Taking expectations on both sides of (13) yields that

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \le \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1} \sigma^{\mathsf{p}} D^{2-\mathsf{p}}, \tag{14}$$

where for the L.H.S., we use $\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \mid \mathcal{F}_{t-1}\right]\right]$ and

$$\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \mid \mathcal{F}_{t-1}\right] = \langle \mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right], \boldsymbol{x}_t - \boldsymbol{x} \rangle = \langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x} \rangle \ge \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}), \tag{15}$$

for the R.H.S., we use $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$.

Finally, we plug $\eta_t = \frac{D}{G\sqrt{t}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}, \forall t \in [T]$ into (14), then use $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \lesssim \sqrt{T}$ and $\sum_{t=1}^{T} \frac{1}{t^{1-1/\mathsf{p}}} \lesssim T^{1/\mathsf{p}}$ to conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

A.2 Extension to Online Strongly Convex Optimization

Next, we extend Theorem 1 to the strongly convex case, i.e., $\exists \mu > 0$ such that for all $t \in [T]$,

$$\frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \langle \nabla \ell_t(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \ell_t(\boldsymbol{y}) \le \ell_t(\boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}, \nabla \ell_t(\boldsymbol{y}) \in \partial \ell_t(\boldsymbol{y}). \tag{16}$$

In this setting, it is well known that OGD achieves a logarithmic regret bound when $\mathsf{p} = 2$ [11, 29]. Theorem 8 below provides the first provable result for $\mathsf{p} < 2$.

Theorem 8. *Under Assumption 1 and additionally assuming (16), taking $\eta_t = \frac{1}{\mu t}$ in OGD (Algorithm 1), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \frac{G^2(1 + \log T)}{\mu} + \frac{\sigma^{\mathsf{p}} G^{2-\mathsf{p}}}{\mu} \times \begin{cases} T^{2-\mathsf{p}} & \mathsf{p} \in (1, 2) \\ 1 + \log T & \mathsf{p} = 2 \end{cases}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Theorem 8 shows that under strongly convexity, OGD for $\mathsf{p} \in (1,2)$ achieves a better sublinear regret $T^{2-\mathsf{p}}$ than $T^{1/\mathsf{p}}$ in Theorem 1 as $2-\mathsf{p} \le 1/\mathsf{p}, \forall \mathsf{p} > 0$. One point we highlight here is that the stepsize $\eta_t = \frac{1}{\mu t}$ is commonly used in the OCO literature and is independent of the tail index p.

However, in contrast to Theorem 1, we suspect Theorem 8 is not tight in T for $\mathsf{p} \in (1,2)$. The reason is that for nonsmooth strongly convex optimization with heavy tails (i.e., $\ell_t = F, \forall t \in [T]$ where F is strongly convex), Theorem 8 can convert to a convergence rate only in the order of $1/T^{\mathsf{p}-1}$, which is worse than the lower bound $1/T^{2-2/\mathsf{p}}$ [49]. Therefore, we conjecture that a way to obtain a better regret bound than $T^{2-\mathsf{p}}$ exists, which we leave as future work.

Proof of Theorem 8. For any $\boldsymbol{x} \in \mathcal{X}$, we take expectations on both sides of (12) to have

$$
\begin{aligned}
\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \le& \left(\frac{1}{\eta_1} - \mu\right)\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \mu\right)\frac{\mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2\right]}{2} \\
&+ \sum_{t=1}^{T} \eta_t G^2 + \mathsf{C}(\mathsf{p})\eta_t^{\mathsf{p}-1}\sigma^{\mathsf{p}} D^{2-\mathsf{p}},
\end{aligned} \tag{17}
$$

where for the L.H.S., we follow a similar step of reasoning out (15) but instead using

$$
\langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x} \rangle \ge \ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}\|^2,
$$

for the R.H.S., we use $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$.

Next, we plug $\eta_t = \frac{1}{\mu t}, \forall t \in [T]$ into (17) to obtain

$$
\begin{aligned}
\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] &\lesssim \sum_{t=1}^{T} \frac{G^2}{\mu t} + \frac{\sigma^{\mathsf{p}} D^{2-\mathsf{p}}}{\mu^{\mathsf{p}-1} t^{\mathsf{p}-1}} \\
&\lesssim \frac{G^2(1+\log T)}{\mu} + \frac{\sigma^{\mathsf{p}} D^{2-\mathsf{p}}}{\mu^{\mathsf{p}-1}} \times \begin{cases} T^{2-\mathsf{p}} & \mathsf{p} \in (1,2) \\ 1+\log T & \mathsf{p} = 2 \end{cases}.
\end{aligned}
$$

Lastly, it is known that if ℓ_t is G-Lipschitz and μ-strongly convex on a domain $\mathcal{X}$ with a diameter D, then it satisfies $D \lesssim \frac{G}{\mu}$ (e.g., see Lemma 2 of [34]). Therefore, when $\mathsf{p} \in (1,2)$,

$$
\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OGD}}(\boldsymbol{x})\right] \lesssim \frac{G^2(1+\log T)}{\mu} + \frac{\sigma^{\mathsf{p}} G^{2-\mathsf{p}}}{\mu} T^{2-\mathsf{p}}.
$$

□

B Missing Proofs for Dual Averaging

This section provides missing proofs for regret bounds of DA.

B.1 Proof of Theorem 2

Proof. Let $L_t(\boldsymbol{x}) \triangleq \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_{t-1}} + \sum_{s=1}^{t-1}\langle \boldsymbol{g}_s, \boldsymbol{x}\rangle, \forall t \in [T+1]$, where $\eta_0 \triangleq \eta_1$. Then DA can be equivalently written as

$$
\boldsymbol{x}_t = \mathrm{argmin}_{\boldsymbol{x} \in \mathcal{X}} L_t(\boldsymbol{x}), \forall t \in [T+1].
$$

By Lemma 7.1 of [29], for any $\boldsymbol{x} \in \mathcal{X}$,

$$
\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle &= \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + L_{T+1}(\boldsymbol{x}_{T+1}) - L_{T+1}(\boldsymbol{x}) + \sum_{t=1}^{T} L_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle - L_{t+1}(\boldsymbol{x}_{t+1}) \\
&\le \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} L_t(\boldsymbol{x}_t) - L_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle,
\end{aligned}
$$

where the inequality holds by $L_{T+1}(\boldsymbol{x}_{T+1}) \le L_{T+1}(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}$ due to $\boldsymbol{x}_{T+1} = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} L_{T+1}(\boldsymbol{x})$. Note that for any $t \in [T]$,

$$\begin{aligned}
&L_t(\boldsymbol{x}_t) - L_{t+1}(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \\
=&L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle + \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_1\|^2}{2\eta_{t-1}} - \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_1\|^2}{2\eta_t} \\
\overset{(a)}{\le}& L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}_{t+1}) + \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \\
\overset{(b)}{\le}& \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}},
\end{aligned}$$

where (a) is by $\eta_t \le \eta_{t-1}, \forall t \in [T]$ and (b) holds because L_t is $\frac{1}{\eta_{t-1}}$-strongly convex and $\boldsymbol{x}_t = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} L_t(\boldsymbol{x})$, which together imply

$$L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}_{t+1}) \le \langle \nabla L_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \le -\frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}}.$$

Therefore, we have

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}}. \tag{18}$$

By the same argument as proving (11) but replacing η_t with η_{t-1}, there is

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \le \eta_{t-1} G^2 + \mathsf{C}(\mathsf{p}) \eta_{t-1}^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}.$$

As such, we know

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \eta_{t-1} G^2 + \mathsf{C}(\mathsf{p}) \eta_{t-1}^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \tag{19}$$

Finally, following similar steps in proving Theorem 1 in Appendix A, we conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

B.2 Dual Averaging with an Adaptive Stepsize

We show that DA with an adaptive stepsize can also achieve the optimal regret $GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}$.

Theorem 9. *Under Assumption 1, taking $\eta_t = 2DV_t^{-1/2}$ and $V_t = \sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2$ in* DA *(Algorithm 2), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. For any $\boldsymbol{x} \in \mathcal{X}$, we have

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \overset{(18)}{\le} \frac{\|\boldsymbol{x} - \boldsymbol{x}_1\|^2}{2\eta_T} + \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}}, \tag{20}$$

where $\eta_0 \triangleq \eta_1$. On the one hand, we can use AM-GM inequality to bound

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \le \frac{\eta_{t-1} \|\boldsymbol{g}_t\|^2}{2}.$$

On the other hand, we know

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} \le \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle \le \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \le \|\boldsymbol{g}_t\| D, \tag{21}$$

where the second step is by Cauchy-Schwarz inequality. Therefore, for any $t \ge 2$,

$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle - \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2}{2\eta_{t-1}} &\le \frac{\eta_{t-1}\|\boldsymbol{g}_t\|^2}{2} \wedge \|\boldsymbol{g}_t\| D \overset{(a)}{\le} \frac{2}{\frac{2}{\eta_{t-1}\|\boldsymbol{g}_t\|^2} + \frac{1}{\|\boldsymbol{g}_t\| D}} \\ &\overset{(b)}{=} \frac{2D\|\boldsymbol{g}_t\|^2}{\sqrt{\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2} + \|\boldsymbol{g}_t\|} \overset{(c)}{\le} \frac{2D\|\boldsymbol{g}_t\|^2}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2}},\end{aligned} \tag{22}$$

where (a) is due to $x \wedge y \le \frac{2}{x^{-1}+y^{-1}}, \forall x, y > 0$, (b) is by $\eta_{t-1} = \frac{2D}{\sqrt{\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2}}$, and (c) holds because of $\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2} \le \sqrt{\sum_{s=1}^{t-1}\|\boldsymbol{g}_s\|^2} + \|\boldsymbol{g}_t\|$. Note that (22) is also true for $t = 1$ by (21).

Combine (20) and (22) and use $\|\boldsymbol{x} - \boldsymbol{x}_1\| \le D$ to obtain

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \le \frac{D^2}{2\eta_T} + \sum_{t=1}^{T} \frac{2D\|\boldsymbol{g}_t\|^2}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2}} = \frac{D^2}{2\eta_T} + \sum_{t=1}^{T} \eta_t \|\boldsymbol{g}_t\|^2,$$

which only differs from (23) by a constant. Hence, by a similar proof for (25), there is

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \lesssim D\left[\sqrt{\sum_{t=1}^{T}\|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \left(\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right],$$

implying

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{DA}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

C Missing Proofs for AdaGrad

This section provides missing proofs for regret bounds of AdaGrad.

C.1 Proof of Theorem 3

Proof. As mentioned, AdaGrad can be viewed as OGD with a stepsize $\eta_t = \frac{\eta}{\sqrt{V_t}} = \frac{\eta}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|^2}}$. Therefore, we can use (1) for AdaGrad to know for any $\boldsymbol{x} \in \mathcal{X}$,

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \le \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \frac{\eta_t\|\boldsymbol{g}_t\|^2}{2}.$$

Sum up the above inequality from $t = 1$ to T and drop the term $-\frac{\|\boldsymbol{x}_{T+1} - \boldsymbol{x}\|^2}{2\eta_T}$ to have

$$\begin{aligned}\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle &\le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}\|^2}{2\eta_1} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T}\frac{\eta_t\|\boldsymbol{g}_t\|^2}{2} \\ &\le \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t\|\boldsymbol{g}_t\|^2}{2},\end{aligned} \tag{23}$$

where the last step is by $\|\boldsymbol{x}_t - \boldsymbol{x}\| \le D, \forall t \in [T]$ and $\eta_{t+1} \le \eta_t, \forall t \in [T-1]$.

Next, observe that for any $t \in [T]$,

$$\|\boldsymbol{g}_t\|^2 = \frac{\eta^2}{\eta_t^2} - \frac{\eta^2}{\eta_{t-1}^2} = \eta^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\left(\frac{1}{\eta_t} + \frac{1}{\eta_{t-1}}\right) \le \frac{2\eta^2}{\eta_t}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right),$$

where $1/\eta_0$ should be read as 0. The above inequality implies

$$\sum_{t=1}^{T} \frac{\eta_t \|\boldsymbol{g}_t\|^2}{2} \le \eta^2 \sum_{t=1}^{T} \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \frac{\eta^2}{\eta_T}. \tag{24}$$

Combine (23) and (24) to have

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \frac{D^2}{2\eta_T} + \frac{\eta^2}{\eta_T} = \left(\frac{D^2}{2\eta} + \eta\right)\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2}.$$

Note that there is

$$\begin{aligned}\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2} &\le \sqrt{\sum_{t=1}^{T} 2\|\nabla \ell_t(\boldsymbol{x}_t)\|^2 + 2\|\boldsymbol{\epsilon}_t\|^2} \le \sqrt{2\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \sqrt{2\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^2} \\ &\le \sqrt{2\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \sqrt{2}\left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},\end{aligned}$$

where the last step is due to $\|\cdot\|_2 \le \|\cdot\|_{\mathsf{p}}$ for any $\mathsf{p} \in [1, 2]$. Hence, we obtain

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \sqrt{2}\left(\frac{D^2}{2\eta} + \eta\right)\left[\sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right]. \tag{25}$$

We take expectations on both sides of (25), then apply Hölder's inequality to have

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \le \left(\sum_{t=1}^{T} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \le \sigma T^{\frac{1}{\mathsf{p}}},$$

and finally plug in $\eta = D/\sqrt{2}$ to conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{AdaGrad}}(\boldsymbol{x})\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}.$$

□

D Missing Proofs for Applications: Nonsmooth Convex Optimization

We prove the following last-iterate convergence result for SGD (i.e., OGD for stochastic optimization) under heavy-tailed noise. The proof of Theorem 10 is inspired by [22, 48].

Theorem 10. *Under Assumption 1 for $\ell_t(\boldsymbol{x}) = F(\boldsymbol{x})$, for any stepsize $\eta_t > 0$ in* OGD *(Algorithm 1), we have*

$$\mathbb{E}[F(\boldsymbol{x}_T) - F(\boldsymbol{x})] \lesssim \frac{D^2}{\sum_{t=1}^{T} \eta_t} + G^2 \sum_{t=1}^{T} \frac{\eta_t^2}{\sum_{s=(t+1)\wedge T}^{T} \eta_s} + \sigma^{\mathsf{p}} D^{2-\mathsf{p}} \sum_{t=1}^{T} \frac{\eta_t^{\mathsf{p}}}{\sum_{s=(t+1)\wedge T}^{T} \eta_s}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. Given $\boldsymbol{x} \in \mathcal{X}$, we recursively define

$$\boldsymbol{y}_0 \triangleq \boldsymbol{x} \quad \text{and} \quad \boldsymbol{y}_t \triangleq \left(1 - \frac{w_{t-1}}{w_t}\right) \boldsymbol{x}_t + \frac{w_{t-1}}{w_t} \boldsymbol{y}_{t-1}, \forall t \in [T], \tag{26}$$

in which

$$w_t \triangleq \frac{\eta_T}{\sum_{s=t+1}^{T} \eta_s}, \forall t \in \{0\} \cup [T-1] \quad \text{and} \quad w_T \triangleq w_{T-1} = 1. \tag{27}$$

Equivalently, $\boldsymbol{y}_t$ can be written into a convex combination of $\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ as

$$\boldsymbol{y}_t = \frac{w_0}{w_t} \boldsymbol{x} + \sum_{s=1}^{t} \frac{w_s - w_{s-1}}{w_t} \boldsymbol{x}_s, \forall t \in \{0\} \cup [T]. \tag{28}$$

Therefore, $\boldsymbol{y}_t$ also falls into $\mathcal{X}$ and satisfies $\boldsymbol{y}_t \in \mathcal{F}_{t-1}$.

We invoke (10) for $\boldsymbol{y}_t$ to obtain

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{y}_t \rangle \leq \frac{\|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|^2}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p}) \eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \tag{29}$$

Since $\boldsymbol{x}_t, \boldsymbol{y}_t \in \mathcal{F}_{t-1}$, there is

$$\mathbb{E}\left[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{y}_t \rangle\right] = \mathbb{E}\left[\langle \mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right], \boldsymbol{x}_t - \boldsymbol{y}_t \rangle\right] = \mathbb{E}\left[\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{y}_t \rangle\right] \geq \mathbb{E}\left[F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right],$$

where the last step is due to the convexity of F. As such, we can take expectations on both sides of (29) to have

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right] &\leq \frac{\mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|^2\right]}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p}) \eta_t^{\mathsf{p}-1} \sigma^{\mathsf{p}} D^{2-\mathsf{p}} \\
&\leq \frac{\mathbb{E}\left[\frac{w_{t-1}}{w_t} \|\boldsymbol{x}_t - \boldsymbol{y}_{t-1}\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|^2\right]}{2\eta_t} + \eta_t G^2 + \mathsf{C}(\mathsf{p}) \eta_t^{\mathsf{p}-1} \sigma^{\mathsf{p}} D^{2-\mathsf{p}},
\end{aligned} \tag{30}$$

where the second step is due to $\|\boldsymbol{x}_t - \boldsymbol{y}_t\|^2 \leq \left(1 - \frac{w_{t-1}}{w_t}\right) \|\boldsymbol{x}_t - \boldsymbol{x}_t\|^2 + \frac{w_{t-1}}{w_t} \|\boldsymbol{x}_t - \boldsymbol{y}_{t-1}\|^2 = \frac{w_{t-1}}{w_t} \|\boldsymbol{x}_t - \boldsymbol{y}_{t-1}\|^2$ by (26) and the convexity of $\|\boldsymbol{x}_t - \cdot\|^2$. Mutiply both sides of (30) by $w_t \eta_t$ and sum up from $t = 1$ to T to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} w_t \eta_t \left(F(\boldsymbol{x}_t) - F(\boldsymbol{y}_t)\right)\right] \\
\leq &\frac{w_0 \|\boldsymbol{x}_1 - \boldsymbol{y}_0\|^2 - \mathbb{E}\left[w_T \|\boldsymbol{x}_{T+1} - \boldsymbol{y}_T\|^2\right]}{2} + \sum_{t=1}^{T} w_t \eta_t^2 G^2 + \mathsf{C}(\mathsf{p}) w_t \eta_t^{\mathsf{p}} \sigma^{\mathsf{p}} D^{2-\mathsf{p}} \\
\leq &\frac{w_0 D^2}{2} + \sum_{t=1}^{T} w_t \eta_t^2 G^2 + \mathsf{C}(\mathsf{p}) w_t \eta_t^{\mathsf{p}} \sigma^{\mathsf{p}} D^{2-\mathsf{p}}.
\end{aligned} \tag{31}$$

Now observe that

$$\begin{aligned}
F(\boldsymbol{y}_t) - F(\boldsymbol{x}) &\overset{(28)}{\leq} \frac{w_0}{w_t} \left(F(\boldsymbol{x}) - F(\boldsymbol{x})\right) + \sum_{s=1}^{t} \frac{w_s - w_{s-1}}{w_t} \left(F(\boldsymbol{x}_s) - F(\boldsymbol{x})\right) \\
&= \sum_{s=1}^{t} \frac{w_s - w_{s-1}}{w_t} \left(F(\boldsymbol{x}_s) - F(\boldsymbol{x})\right),
\end{aligned}$$

which implies

$$\sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{y}_t)-F(\boldsymbol{x})\right) \le \sum_{t=1}^{T}\sum_{s=1}^{t}\left(w_s-w_{s-1}\right)\eta_t\left(F(\boldsymbol{x}_s)-F(\boldsymbol{x})\right)$$
$$= \sum_{t=1}^{T}\left(w_t-w_{t-1}\right)\left(\sum_{s=t}^{T}\eta_s\right)\left(F(\boldsymbol{x}_t)-F(\boldsymbol{x})\right).$$

Thus, we can lower bound the L.H.S. of (31) by

$$\begin{aligned}\sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{x}_t)-F(\boldsymbol{y}_t)\right) &= \sum_{t=1}^{T} w_t\eta_t\left(F(\boldsymbol{x}_t)-F(\boldsymbol{x})\right)-w_t\eta_t\left(F(\boldsymbol{y}_t)-F(\boldsymbol{x})\right)\\ &\ge \sum_{t=1}^{T}\left[w_t\eta_t-\left(w_t-w_{t-1}\right)\left(\sum_{s=t}^{T}\eta_s\right)\right]\left(F(\boldsymbol{x}_t)-F(\boldsymbol{x})\right)\\ &= w_T\eta_T\left(F(\boldsymbol{x}_T)-F(\boldsymbol{x})\right), \end{aligned} \tag{32}$$

where the last step is due to, for $t\in[T-1]$,

$$\begin{aligned} w_t\eta_t-\left(w_t-w_{t-1}\right)\left(\sum_{s=t}^{T}\eta_s\right) &\overset{(27)}{=} \frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}\cdot\eta_t-\left(\frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}-\frac{\eta_T}{\sum_{s=t}^{T}\eta_s}\right)\left(\sum_{s=t}^{T}\eta_s\right)\\ &= \frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}\cdot\eta_t-\frac{\eta_T}{\sum_{s=t+1}^{T}\eta_s}\cdot\eta_t=0, \end{aligned}$$

and $w_T\overset{(27)}{=}w_{T-1}=1$.

We plug (32) back into (31) and divide both sides by $w_T\eta_T$to obtain

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{x}_T)-F(\boldsymbol{x})\right] &\le \frac{w_0D^2}{2w_T\eta_T}+\sum_{t=1}^{T}\frac{w_t\eta_t^2}{w_T\eta_T}G^2+\mathsf{C}(\mathsf{p})\frac{w_t\eta_t^{\mathsf{p}}}{w_T\eta_T}\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\\ &\overset{(27)}{\lesssim} \frac{D^2}{\sum_{t=1}^{T}\eta_t}+G^2\sum_{t=1}^{T}\frac{\eta_t^2}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}+\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\sum_{t=1}^{T}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}. \end{aligned}$$

□

Equipped with Theorem 10, we show the following anytime last-iterate convergence rate for SGD/OGD. As far as we know, this is the first and the only provable result demonstrating that the last iterate of SGD can converge in heavy-tailed stochastic optimization without gradient clipping. Compared to Corollary 2, the difference is up to an extra logarithmic factor. Therefore, it is nearly optimal.

Corollary 5. *Under Assumption 1 for $\ell_t(\boldsymbol{x})=F(\boldsymbol{x})$, taking $\eta_t=\frac{D}{G\sqrt{t}}\wedge\frac{D}{\sigma t^{1/\mathsf{p}}}$ in* OGD *(Algorithm 1), we have*

$$\mathbb{E}\left[F(\boldsymbol{x}_T)-F(\boldsymbol{x})\right]\lesssim\frac{GD\left(1+\log T\right)}{\sqrt{T}}+\frac{\sigma D\left(1+\log T\right)}{T^{1-\frac{1}{\mathsf{p}}}},\forall\boldsymbol{x}\in\mathcal{X}.$$

Proof. By Theorem 10, we have

$$\begin{aligned}&\mathbb{E}\left[F(\boldsymbol{x}_T)-F(\boldsymbol{x})\right]\\ \lesssim&\frac{D^2}{\sum_{t=1}^{T}\eta_t}+G^2\sum_{t=1}^{T}\frac{\eta_t^2}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}+\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\sum_{t=1}^{T}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=(t+1)\wedge T}^{T}\eta_s}\\ =&\frac{D^2}{\sum_{t=1}^{T}\eta_t}+G^2\left(\eta_T+\sum_{t=1}^{T-1}\frac{\eta_t^2}{\sum_{s=t+1}^{T}\eta_s}\right)+\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\left(\eta_T^{\mathsf{p}-1}+\sum_{t=1}^{T-1}\frac{\eta_t^{\mathsf{p}}}{\sum_{s=t+1}^{T}\eta_s}\right).\end{aligned}$$

For any $t \in \{0\} \cup [T-1]$, observe that by Cauchy-Schwarz inequality

$$(T-t)^2 \le \left(\sum_{s=t+1}^{T} \frac{1}{\eta_s}\right)\left(\sum_{s=t+1}^{T} \eta_s\right) \Rightarrow \frac{1}{\sum_{s=t+1}^{T} \eta_s} \le \frac{\sum_{s=t+1}^{T} \frac{1}{\eta_s}}{(T-t)^2}.$$

Thus, there is

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{x}_T) - F(\boldsymbol{x})\right] \lesssim & \frac{D^2}{T^2}\sum_{t=1}^{T}\frac{1}{\eta_t} + G^2\left(\eta_T + \sum_{t=1}^{T-1}\frac{\eta_t^2 \sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right) \\ & + \sigma^{\mathsf{p}} D^{2-\mathsf{p}}\left(\eta_T^{\mathsf{p}-1} + \sum_{t=1}^{T-1}\frac{\eta_t^{\mathsf{p}}\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right). \end{aligned} \tag{33}$$

We first bound

$$\sum_{t=1}^{T}\frac{1}{\eta_t} = \sum_{t=1}^{T}\frac{G\sqrt{t}}{D} \vee \frac{\sigma t^{1/\mathsf{p}}}{D} \le \sum_{t=1}^{T}\frac{G\sqrt{t}}{D} + \frac{\sigma t^{1/\mathsf{p}}}{D} \lesssim \frac{G}{D}T^{3/2} + \frac{\sigma}{D}T^{1+1/\mathsf{p}},$$

which implies

$$\frac{D^2}{T^2}\sum_{t=1}^{T}\frac{1}{\eta_t} \lesssim \frac{GD}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{34}$$

Next, we know

$$\begin{aligned}\eta_T + \sum_{t=1}^{T-1}\frac{\eta_t^2\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2} &\overset{(a)}{\le} \frac{D}{G\sqrt{T}} + \sum_{t=1}^{T-1}\left[\frac{D}{G}\cdot\frac{\sum_{s=t+1}^{T}\sqrt{s}}{t(T-t)^2} + \frac{\sigma D}{G^2}\cdot\frac{\sum_{s=t+1}^{T}s^{1/\mathsf{p}}}{t(T-t)^2}\right] \\ &\overset{\text{Fact 1}}{\lesssim} \frac{D}{G\sqrt{T}} + \frac{D\left(1+\log T\right)}{G\sqrt{T}} + \frac{\sigma D\left(1+\log T\right)}{G^2 T^{1-\frac{1}{\mathsf{p}}}},\end{aligned}$$

where (a) is by $\eta_t \le \frac{D}{G\sqrt{t}}$ and $\frac{1}{\eta_s} \le \frac{G\sqrt{s}}{D} \vee \frac{\sigma s^{1/\mathsf{p}}}{D}$. Hence, there is

$$G^2\left(\eta_T + \sum_{t=1}^{T-1}\frac{\eta_t^2\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right) \lesssim \frac{GD\left(1+\log T\right)}{\sqrt{T}} + \frac{\sigma D\left(1+\log T\right)}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{35}$$

Similarly, we can bound

$$\sigma^{\mathsf{p}}D^{2-\mathsf{p}}\left(\eta_T^{\mathsf{p}-1} + \sum_{t=1}^{T-1}\frac{\eta_t^{\mathsf{p}}\sum_{s=t+1}^{T}\frac{1}{\eta_s}}{(T-t)^2}\right) \lesssim \frac{GD\left(1+\log T\right)}{\sqrt{T}} + \frac{\sigma D\left(1+\log T\right)}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{36}$$

Finally, we plug (34), (35) and (36) back into (33) to conclude. □

E Missing Proofs for Applications: Nonsmooth Nonconvex Optimization

E.1 (δ, ϵ)-Stationary Points

Definition 2 (Definition 4 of [7])**.** A point $\boldsymbol{x} \in \mathbb{R}^d$ is a (δ, ϵ)-stationary point of an almost-everywhere differentiable function F if there is a finite subset $S \subset \mathcal{B}^d(\boldsymbol{x}, \delta)$ such that for $\boldsymbol{y}$ selected uniformly at random from S, $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$ and $\|\mathbb{E}[\nabla F(\boldsymbol{y})]\| \le \epsilon$.

The concept of the (δ, ϵ)-stationary point presented here is due to [7], which is mildly more stringent than the notion of [50], since the latter does not require $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$. For more discussions, see Section 2.1 of [7].

E.2 Proof of Theorem 4

In this section, our ultimate goal is to prove Theorem 4 for the O2NC algorithm, extending Theorem 8 of [7] from $\mathsf{p}=2$ to any $\mathsf{p}\in(1,2]$. Notably, our new result does not require any modification to the O2NC method, but is obtained only from a more careful analysis, indicating that O2NC is a robust and powerful algorithmic framework.

We begin with Lemma 2, which lies as the cornerstone for establishing the convergence of O2NC.

Lemma 2 (Theorem 7 of [7]). *Under Assumption 2 (only need the second point and the unbiased part in the fourth point), for any sequence of vectors $\boldsymbol{u}_1,\dots,\boldsymbol{u}_{KT}\in\mathbb{R}^d$, O2NC (Algorithm 4) guarantees*

$$\mathbb{E}\left[F(\boldsymbol{y}_{KT})\right]=F(\boldsymbol{y}_0)+\mathbb{E}\left[\sum_{n=1}^{KT}\langle\boldsymbol{g}_n,\boldsymbol{x}_n-\boldsymbol{u}_n\rangle\right]+\mathbb{E}\left[\sum_{n=1}^{KT}\langle\boldsymbol{g}_n,\boldsymbol{u}_n\rangle\right]. \tag{37}$$

To relate Lemma 2 to the concept of K-shifting regret introduced before (see (9)), suppose now a sequence of vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_K$ is given, if we set $\boldsymbol{u}_n=\boldsymbol{v}_k$ for all $n\in\{(k-1)T+1,\dots,kT\}$ and $k\in[K]$, then the second term on the R.H.S. of (37) can be written as $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_K)\right]$, and the third term can be simplified into $\sum_{k=1}^{K}\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right]$.

Same as [7], we pick $\boldsymbol{v}_k\triangleq -D\frac{\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)}{\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|}$ for some constant $D>0$, which gives us

$$\begin{aligned}\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right]&=\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n,\boldsymbol{v}_k\right\rangle\right]-D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right]\\&\le D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\right]-D\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\nabla F(\boldsymbol{z}_n)\right\|\right].\end{aligned}$$

If $\boldsymbol{\epsilon}_n$ has a finite variance (i.e., $\mathsf{p}=2$), then like [7], one can invoke Hölder's inequality and use the fact $\mathbb{E}\left[\langle\boldsymbol{\epsilon}_m,\boldsymbol{\epsilon}_n\rangle\right]=0,\forall m\neq n\in[KT]$ to obtain for any $k\in[K]$,

$$\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|\right]\le\sqrt{\mathbb{E}\left[\left\|\sum_{n=(k-1)T+1}^{kT}\boldsymbol{\epsilon}_n\right\|^2\right]}=\sqrt{\sum_{n=(k-1)T+1}^{kT}\mathbb{E}\left[\|\boldsymbol{\epsilon}_n\|^2\right]}\le\sigma\sqrt{T}.$$

However, this argument immediately fails when $\mathsf{p}<2$ as $\mathbb{E}\left[\|\boldsymbol{\epsilon}_n\|^2\right]$ can be $+\infty$. To handle this potential issue, we require the following Lemma 3.

Lemma 3 (Lemma 4.3 of [23]). *Given a vector-valued martingale difference sequence $\boldsymbol{w}_1,\dots,\boldsymbol{w}_T$, there is*

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T}\boldsymbol{w}_t\right\|\right]\le 2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{w}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right],\forall\mathsf{p}\in[1,2].$$

Equipped with Lemmas 2 and 3, we are ready to formally prove Theorem 4, demonstrating that the O2NC framework provably works under heavy-tailed noise.

Proof of Theorem 4. We invoke Lemma 2 with $\boldsymbol{u}_n=\boldsymbol{v}_{\lceil n/T\rceil},\forall n\in[KT]$ (equivalently, $\boldsymbol{u}_n=\boldsymbol{v}_k$ if $n\in\{(k-1)T+1,\dots,kT\}$) and use the definition of K-shifting regret (see (9)) to obtain

$$\mathbb{E}\left[F(\boldsymbol{y}_{KT})\right]=F(\boldsymbol{y}_0)+\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}\left(\boldsymbol{v}_1,\dots,\boldsymbol{v}_K\right)\right]+\sum_{k=1}^{K}\mathbb{E}\left[\left\langle\sum_{n=(k-1)T+1}^{kT}\boldsymbol{g}_n,\boldsymbol{v}_k\right\rangle\right]. \tag{38}$$

Recall that $\boldsymbol{g}_n = \nabla F(\boldsymbol{z}_n) + \boldsymbol{\epsilon}_n$, which implies for any $k \in [K]$,

$$\begin{aligned}\mathbb{E}\left[\left\langle \sum_{n=(k-1)T+1}^{kT} \boldsymbol{g}_n, \boldsymbol{v}_k \right\rangle\right] &= \mathbb{E}\left[\left\langle \sum_{n=(k-1)T+1}^{kT} \boldsymbol{\epsilon}_n, \boldsymbol{v}_k \right\rangle\right] + \mathbb{E}\left[\left\langle \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n), \boldsymbol{v}_k \right\rangle\right] \\ &\le \mathbb{E}\left[\left\| \sum_{n=(k-1)T+1}^{kT} \boldsymbol{\epsilon}_n \right\| \|\boldsymbol{v}_k\|\right] + \mathbb{E}\left[\left\langle \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n), \boldsymbol{v}_k \right\rangle\right] \\ &= D\mathbb{E}\left[\left\| \sum_{n=(k-1)T+1}^{kT} \boldsymbol{\epsilon}_n \right\|\right] - D\mathbb{E}\left[\left\| \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n) \right\|\right], \end{aligned} \tag{39}$$

where the second step is by Cauchy-Schwarz inequality and the last equation holds due to

$$\boldsymbol{v}_k = -D \frac{\sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n)}{\left\| \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n) \right\|}, \forall k \in [K]\,. \tag{40}$$

Combine (38) and (39), apply $F(\boldsymbol{y}_{KT}) \ge F_\star$, and rearrange terms to have

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{K} \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n) \right\|\right] \\ \le &\frac{F(\boldsymbol{y}_0) - F_\star}{DKT} + \frac{\mathbb{E}\left[R_T^{\mathsf{A}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)\right]}{DKT} + \frac{\sum_{k=1}^{K} \mathbb{E}\left[\left\| \sum_{n=(k-1)T+1}^{kT} \boldsymbol{\epsilon}_n \right\|\right]}{KT}.\end{aligned} \tag{41}$$

For any fixed $k \in [K]$, we apply Lemma 3 with $\boldsymbol{w}_t = \boldsymbol{\epsilon}_{(k-1)T+t}, \forall t \in [T]$ to know

$$\begin{aligned}\mathbb{E}\left[\left\| \sum_{n=(k-1)T+1}^{kT} \boldsymbol{\epsilon}_n \right\|\right] &\le 2\sqrt{2}\mathbb{E}\left[\left(\sum_{n=(k-1)T+1}^{kT} \|\boldsymbol{\epsilon}_n\|^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}\right] \\ &\le 2\sqrt{2}\left(\sum_{n=(k-1)T+1}^{kT} \mathbb{E}\left[\|\boldsymbol{\epsilon}_n\|^{\mathsf{p}}\right] \right)^{\frac{1}{\mathsf{p}}} \le 2\sqrt{2}\sigma T^{\frac{1}{\mathsf{p}}},\end{aligned} \tag{42}$$

where the second step is by Hölder's inequality (note that $\mathsf{p} > 1$). Finally, we conclude the proof after plugging (42) back into (41). □

E.3 Proof of Theorem 5

Proof. By Theorem 4, there is

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{K} \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n) \right\|\right] \lesssim \frac{F(\boldsymbol{y}_0) - F_\star}{DKT} + \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)\right]}{DKT} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{43}$$

Note that A has the domain $\mathcal{X} = \mathcal{B}^d(D)$ and $s_n \sim \mathsf{Uniform}\,[0,1]$. Thus, for any $n \in [KT]$,

$$\|\boldsymbol{x}_n\| \le D \quad \text{and} \quad s_n \in [0,1]\,. \tag{44}$$

We first lower bound the L.H.S. of (43). Given $k \in [K]$, for any $m < n \in \{(k-1)T+1, \ldots, kT\}$, observe that

$$\begin{aligned}\|\boldsymbol{z}_n - \boldsymbol{z}_m\| = \left\|\boldsymbol{y}_{n-1} + s_n \boldsymbol{x}_n - \boldsymbol{y}_{m-1} - s_m \boldsymbol{x}_m\right\| = \left\| s_n \boldsymbol{x}_n - s_m \boldsymbol{x}_m + \sum_{i=m}^{n-1} \boldsymbol{x}_i \right\| \\ \le s_n \|\boldsymbol{x}_n\| + (1 - s_m)\|\boldsymbol{x}_m\| + \sum_{i=m+1}^{n-1} \|\boldsymbol{x}_i\| \overset{(44)}{\le} (n - m + 1)\,D \le DT.\end{aligned}$$

Recall that $\bar{\boldsymbol{z}}_k = \frac{1}{T}\sum_{n=(k-1)T+1}^{kT} \boldsymbol{z}_n$ and $D = \delta/T$ now, then the above inequality implies

$$\|\boldsymbol{z}_n - \bar{\boldsymbol{z}}_k\| \le DT = \delta, \forall n \in \{(k-1)T+1, \dots, kT\}, \tag{45}$$

which means

$$\boldsymbol{z}_n \in \mathcal{B}^d(\bar{\boldsymbol{z}}_k, \delta), \forall n \in \{(k-1)T+1, \dots, kT\}.$$

By the definition of $\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta$ (see Definition 1), there is

$$\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta \le \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n) \right\|. \tag{46}$$

Next, we upper bound the R.H.S. of (43). By the definition of K-shifting regret (see (9)), there is

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_K)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\sum_{n=(k-1)T+1}^{kT} \langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k \rangle\right].$$

Note that we reset the stepsize in A after every T iterations and $\boldsymbol{v}_k \in \mathcal{B}^d(D)$ by its definition (see (40)). Then for any $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$, we can invoke its regret bound[2] (i.e., Theorems 1, 2 and 3) to obtain

$$\mathbb{E}\left[\sum_{n=(k-1)T+1}^{kT} \langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k \rangle\right] \lesssim GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}, \forall k \in [K],$$

which implies

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_K)\right] \lesssim GDK\sqrt{T} + \sigma DKT^{1/\mathsf{p}}. \tag{47}$$

Finally, we plug (46) and (47) back into (43), then use $D = \delta/T$ and $\Delta = F(\boldsymbol{y}_0) - F_\star$ to have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\Delta}{\delta K} + \frac{G}{\sqrt{T}} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

□

E.4 Proof of Corollary 3

Proof. Recall that we pick

$$K = \left\lfloor \frac{N}{T} \right\rfloor \quad \text{and} \quad T = \left\lceil \frac{N}{2} \right\rceil \wedge \left(\left\lceil \left(\frac{\delta G N}{\Delta}\right)^{\frac{2}{3}} \right\rceil \vee \left\lceil \left(\frac{\delta \sigma N}{\Delta}\right)^{\frac{\mathsf{p}}{2\mathsf{p}-1}} \right\rceil \right),$$

where $\Delta = F(\boldsymbol{y}_0) - F_\star$. We invoke Theorem 5 and use $KT \ge N/4$ (see Fact 2) to obtain

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\Delta T}{\delta N} + \frac{G}{\sqrt{T}} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

By the definition of T, we know

$$\frac{\Delta T}{\delta N} \lesssim \frac{\Delta}{\delta N}\left[1 + \left(\frac{\delta G N}{\Delta}\right)^{\frac{2}{3}} + \left(\frac{\delta \sigma N}{\Delta}\right)^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\right] = \frac{\Delta}{\delta N} + \frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}} + \frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}},$$

[2] A minor point here is that the current function $\ell_n(\boldsymbol{x}) = \langle \boldsymbol{g}_n, \boldsymbol{x} \rangle$ does not entirely fit Assumption 1. We clarify that one does not need to worry about it, since all results proved in Section 3 hold under this change. For example, in the proof of Theorem 1, we can safely replace the L.H.S. of (14) with $\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle\right]$.

and

$$\frac{G}{\sqrt{T}} \lesssim \frac{G}{\sqrt{N}} + \frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}}, \qquad \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}} \lesssim \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

Therefore, there is

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{G}{\sqrt{N}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \frac{\Delta}{\delta N} + \frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}} + \frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

□

E.5 Extension to the Case of Unknown Problem-Dependent Parameters

In Corollary 6, we show how to set K and T when all problem-dependent parameters are unknown. It is particularly meaningful for AdaGrad. As in that case, the rate is achieved without knowing any problem-dependent parameter. This kind of result is the first to appear for nonsmooth nonconvex optimization with heavy tails. However, the rate is not as good as Corollary 3. It is currently unclear whether the same bound $1/(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}$ as in Corollary 3 can be obtained when no information about the problem is known.

Corollary 6. *Under the same setting of Theorem 5, suppose we have $N \geq 2$ stochastic gradient budgets, taking $K = \lfloor N/T \rfloor$ and $T = \lceil N/2 \rceil \wedge \left\lceil (\delta N)^{\frac{2}{3}} \right\rceil$, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\Delta}{(\delta N)\wedge(\delta N)^{\frac{1}{3}}} + \frac{G}{\sqrt{N}\wedge(\delta N)^{\frac{1}{3}}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}\wedge(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

Proof. We invoke Theorem 5 and use $KT \geq N/4$ (see Fact 2) to obtain

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\Delta T}{\delta N} + \frac{G}{\sqrt{T}} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

By the definition of T, we know

$$\frac{\Delta T}{\delta N} \lesssim \frac{\Delta}{\delta N}\left[1 + (\delta N)^{\frac{2}{3}}\right] \lesssim \frac{\Delta}{(\delta N)\wedge(\delta N)^{\frac{1}{3}}}.$$

and

$$\frac{G}{\sqrt{T}} \lesssim \frac{G}{\sqrt{N}} + \frac{G}{(\delta N)^{\frac{1}{3}}} \lesssim \frac{G}{\sqrt{N}\wedge(\delta N)^{\frac{1}{3}}},$$
$$\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}} \lesssim \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \frac{\sigma}{(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}} \lesssim \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}\wedge(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

Therefore, there is

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\Delta}{(\delta N)\wedge(\delta N)^{\frac{1}{3}}} + \frac{G}{\sqrt{N}\wedge(\delta N)^{\frac{1}{3}}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}\wedge(\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

□

E.6 Proof of Theorem 6

In this section, our ultimate goal is to prove Theorem 6. The analysis follows the framework first established in [3] and later developed by [7] but with some necessary (though minor) variation to make it compatible with heavy-tailed noise. In the following, we will slightly abuse the notation A to denote any possible randomized first-order algorithm instead of an online learning algorithm.

E.6.1 Basic Definitions

To begin with, we introduce some basic definitions given in [3].

Definition 3 (stochastic first-order oracle)**.** Given a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, a tuple $(\boldsymbol{g}, \mathcal{R}, \mathbb{P}_r)$ is called a stochastic first-order oracle of F if $\mathbb{P}_r$ is a probability distribution on the measurable space $\mathcal{R}$ and $\boldsymbol{g} : \mathbb{R}^d \times \mathcal{R} \to \mathbb{R}^d$ satisfies $\mathbb{E}_{r\sim\mathbb{P}_r}[\boldsymbol{g}(\boldsymbol{x}, r)] = \nabla F(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^d$.

Remark 9. When the context is clear, we will omit F, $\mathcal{R}$ and $\mathbb{P}_r$, and simply call $\boldsymbol{g}$ as a stochastic first-order oracle.

Definition 4 (randomized algorithm)**.** A randomized algorithm A consists of a probability distribution $\mathbb{P}_s$ over a measurable space $\mathcal{S}$ and a sequence of measurable mappings $\mathsf{A}_t, \forall t \in \mathbb{N}$ such that every A_t takes a common random seed $s \in \mathcal{S}$ and the first $t-1$ oracle responses to produce the t-th query. Concretely, given a differentiable function F equipped with a stochastic first-order oracle $(\boldsymbol{g}, \mathcal{R}, \mathbb{P}_r)$, the sequence $\boldsymbol{x}_t, \forall t \in \mathbb{N}$ produced by A to optimize F is recursively defined as

$$\boldsymbol{x}_t = \mathsf{A}_t(s, \boldsymbol{g}(\boldsymbol{x}_{t-1}, r_{t-1}), \ldots, \boldsymbol{g}(\boldsymbol{x}_1, r_1)), \forall t \in \mathbb{N},$$

where $s \sim \mathbb{P}_s$ is drawn a single time at the beginning of the algorithm and $r_t \sim \mathbb{P}_r, \forall t \in \mathbb{N}$ is a sequence of i.i.d. random variables. Moreover, $\mathsf{A}_{\mathsf{rand}}$ denotes the set containing all randomized algorithms.

Next, we require a useful concept named probability-q zero-chain. Before formally stating what it is, we need some notations. Given $\boldsymbol{x} \in \mathbb{R}^d$ and $\alpha \in [0, 1]$, $\mathsf{prog}_\alpha(\boldsymbol{x})$ denotes the largest index whose entry is α-far from 0, i.e.,

$$\mathsf{prog}_\alpha(\boldsymbol{x}) \triangleq \max\{i \in [d] : |\boldsymbol{x}[i]| > \alpha\} \quad \text{where} \quad \max\varnothing \triangleq 0.$$

In addition, for any $j \in \{0\} \cup \mathbb{N}$, let $\boldsymbol{x}_{\leq j}[i] \triangleq \boldsymbol{x}[i]\,\mathbb{1}[i \leq j], \forall i \in [d]$ be the truncated version of $\boldsymbol{x}$. Now we are ready to provide the definition of the probability-q zero-chain.

Definition 5 (probability-q zero-chain)**.** A stochastic first-order oracle $(\boldsymbol{g}, \mathcal{R}, \mathbb{P}_r)$ is called a probability-q zero-chain if and only if

$$\mathbb{P}_r\left[\forall \boldsymbol{x} \in \mathbb{R}^d : \mathsf{prog}_0(\boldsymbol{g}(\boldsymbol{x}, r)) \leq \mathsf{prog}_{1/4}(\boldsymbol{x}) \text{ and } \boldsymbol{g}(\boldsymbol{x}, r) = \boldsymbol{g}(\boldsymbol{x}_{\leq \mathsf{prog}_{1/4}(\boldsymbol{x})}, r)\right] \geq 1 - q,$$

and

$$\mathbb{P}_r\left[\forall \boldsymbol{x} \in \mathbb{R}^d : \mathsf{prog}_0(\boldsymbol{g}(\boldsymbol{x}, r)) \leq \mathsf{prog}_{1/4}(\boldsymbol{x}) + 1 \text{ and } \boldsymbol{g}(\boldsymbol{x}, r) = \boldsymbol{g}(\boldsymbol{x}_{\leq \mathsf{prog}_{1/4}(\boldsymbol{x})+1}, r)\right] = 1.$$

E.6.2 Useful Existing Results

In this part, we list some useful existing results from [3] and [7].

Given $d \geq T \in \mathbb{N}$ and a differentiable function $F_T : \mathbb{R}^T \to \mathbb{R}$ with a stochastic first-order oracle $\boldsymbol{g}_T$ that is a probability-q zero-chain, their rotated variants parametrized by a matrix $U \in \mathsf{Ortho}(d, T) \triangleq \{U \in \mathbb{R}^{d\times T} : U^\top U = I_T\}$ (where I_T is the T-dimensional identity matrix) are defined as

$$\bar{F}_{T,U}(\boldsymbol{x}) \triangleq F_T(U^\top \boldsymbol{x}) \quad \text{and} \quad \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r) \triangleq U\boldsymbol{g}_T(U^\top \boldsymbol{x}, r). \tag{48}$$

Clearly, $\bar{\boldsymbol{g}}_{T,U}$ is a stochastic first-order oracle of $\bar{F}_{T,U}$. In addition, we emphasize that the input variable $\boldsymbol{x}$ is from $\mathbb{R}^d$ instead of $\mathbb{R}^T$ now.

With $\bar{F}_{T,U}$ and $\bar{\boldsymbol{g}}_{T,U}$, we state the following lemma due to [3].

Lemma 4 (Lemma 5 of [3])**.** *Given $q, \iota \in (0, 1)$, $R > 0$, $T \in \mathbb{N}$, $d \in \mathbb{N}$ satisfying $d \geq T + 32R^2 \log \frac{2T^2}{q\iota}$ and a randomized algorithm $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$, suppose the output of A always has a norm bounded by R, let $\boldsymbol{x}_t, \forall t \in \mathbb{N}$ be the trajectory produced by applying A to optimize $\bar{F}_{T,U}$ interacting with the stochastic first-order oracle $\bar{\boldsymbol{g}}_{T,U}$, where U is drawn from $\mathsf{Ortho}(d, T)$ uniformly, then there is*

$$\Pr\left[\mathsf{prog}_{1/4}\left(U^\top \boldsymbol{x}_t\right) < T, \forall t \leq \frac{T - \log(2/\iota)}{2q}\right] \geq 1 - \iota.$$

Here Pr *takes into account all randomness over A, $\boldsymbol{g}_T$, and U.*

Note that Lemma 4 can be only applied to a randomzied algorithm with bounded outputs. To overcome this issue, we need the following variants of $\bar{F}_{T,U}$ and $\bar{\boldsymbol{g}}_{T,U}$ introduced by [7]:

$$\widehat{F}_{T,U}(\boldsymbol{x}) \triangleq \bar{F}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta \boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), \tag{49}$$

$$\widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r) \triangleq \mathcal{J}_{R,d}(\boldsymbol{x})^\top \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r) + \eta \nabla(\boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})), \tag{50}$$

where $\boldsymbol{\rho}_{R,d}(\boldsymbol{x}) \triangleq \frac{\boldsymbol{x}}{\sqrt{1+\|\boldsymbol{x}\|^2/R^2}}$ is a bijection from $\mathbb{R}^d$ to $\mathsf{int}\mathcal{B}^d(R)$[3], $\mathcal{J}_{R,d}$ denotes the Jacobian of $\boldsymbol{\rho}_{R,d}$, and $R, \eta > 0$ are two constants being determined later.

Before moving on, we state some useful properties of $\boldsymbol{\rho}_{R,d}$ here.

Lemma 5 (Lemma 13 of [3] and Proposition 29 of [7])**.** *Let $\|\cdot\|_{\mathsf{op}}$ denotes the operator norm, then for $\boldsymbol{\rho}_{R,d} : \mathbb{R}^d \to \mathbb{R}^d, \boldsymbol{x} \mapsto \frac{\boldsymbol{x}}{\sqrt{1+\|\boldsymbol{x}\|^2/R^2}}$, there are*

1. $\nabla(\boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) = \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x})$,
2. $\left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x}) - \boldsymbol{\rho}_{R,d}(\boldsymbol{y})\right\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$,
3. $\mathcal{J}_{R,d}(\boldsymbol{x}) = \frac{I_d - \boldsymbol{\rho}_{R,d}(\boldsymbol{x})\boldsymbol{\rho}_{R,d}(\boldsymbol{x})^\top / R^2}{\sqrt{1+\|\boldsymbol{x}\|^2/R^2}}$,
4. $\|\mathcal{J}_{R,d}(\boldsymbol{x})\|_{\mathsf{op}} = \frac{1}{\sqrt{1+\|\boldsymbol{x}\|^2/R^2}} \leq 1$,
5. $\|\mathcal{J}_{R,d}(\boldsymbol{x}) - \mathcal{J}_{R,d}(\boldsymbol{y})\|_{\mathsf{op}} \leq \frac{3}{R} \|\boldsymbol{x} - \boldsymbol{y}\|$.

We then can combine (48), (49), (50) and Lemma 5 to have

$$\widehat{F}_{T,U}(\boldsymbol{x}) = \bar{F}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta \boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}) = F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta \boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), \tag{51}$$

$$\begin{aligned}
\nabla \widehat{F}_{T,U}(\boldsymbol{x}) &= \mathcal{J}_{R,d}(\boldsymbol{x})^\top \nabla \bar{F}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x}) & (52)\\
&= \mathcal{J}_{R,d}(\boldsymbol{x})^\top U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), & (53)\\
\widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r) &= \mathcal{J}_{R,d}(\boldsymbol{x})^\top \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r) + \eta \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x}) & (54)\\
&= \mathcal{J}_{R,d}(\boldsymbol{x})^\top U \boldsymbol{g}_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r) + \eta \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x}). & (55)
\end{aligned}$$

Now we are ready to give the following Lemma 6, which is almost identical to Lemma 25 in [7] by differing up to numerical constants. Since Lemma 6 is particularly important, we therefore provide its full proof here.

Lemma 6 (Lemma 25 of [7])**.** *For $T \in \mathbb{N}$, suppose that F_T satisfies the following two requirements:*

1. *For all $\boldsymbol{x} \in \mathbb{R}^T$, $\|\nabla F_T(\boldsymbol{x})\| \leq \gamma\sqrt{T}$ where $\gamma > 0$ is a constant satisfying $\frac{1}{\sqrt{5}} - \frac{39}{140} \geq \frac{10}{63} + \frac{5\sqrt{5}}{126\gamma} \Leftrightarrow \gamma \geq \frac{50\sqrt{5}}{252\sqrt{5}-551} \approx 8.95$.*
2. *For all $\boldsymbol{x} \in \mathbb{R}^T$, if $\mathsf{prog}_1(\boldsymbol{x}) < T$, then $\|\nabla F_T(\boldsymbol{x})\| \geq |\nabla F_T(\boldsymbol{x})[\mathsf{prog}_1(\boldsymbol{x}) + 1]| > 1$.*

Given $q, \iota \in (0,1)$, $\eta = \frac{1}{\sqrt{5}} - \frac{39}{140}$, $R = 7\gamma\sqrt{T}$, $d \in \mathbb{N}$ satisfying $d \geq T + 32R^2 \log \frac{2T^2}{q\iota}$ and a randomized algorithm $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$, let $\boldsymbol{x}_t, \forall t \in \mathbb{N}$ be the trajectory produced by applying A to optimize $\widehat{F}_{T,U}$ interacting with the stochastic first-order oracle $\widehat{\boldsymbol{g}}_{T,U}$, where U is drawn from $\mathsf{Ortho}(d,T)$ uniformly, then there is

$$\Pr\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\| \geq \frac{1}{2}, \forall t \leq \frac{T - \log(2/\iota)}{2q}\right] \geq 1 - \iota.$$

Here $\Pr$ *takes into account all randomness over* A, $\boldsymbol{g}_T$ *and* U.

[3] As one can check, $\boldsymbol{\rho}_{R,d}^{-1}(\boldsymbol{x}) : \mathsf{int}\mathcal{B}^d(R) \to \mathbb{R}^d$ exists and equals $\frac{\boldsymbol{x}}{\sqrt{1-\|\boldsymbol{x}\|^2/R^2}}$.

Proof. By definition of randomized algorithms (see Definition 4), we have

$$\boldsymbol{x}_t = \mathsf{A}_t\left(s, \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}_{t-1}, r_{t-1}), \ldots, \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}_1, r_1)\right), \forall t \in \mathbb{N}. \tag{56}$$

Now let us consider another sequence

$$\boldsymbol{y}_t \triangleq \boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t) = \frac{\boldsymbol{x}_t}{\sqrt{1 + \|\boldsymbol{x}_t\|/R^2}} \in \mathsf{int}\mathcal{B}^d(R), \forall t \in \mathbb{N}. \tag{57}$$

Note that

$$\begin{aligned}
\widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}_t, r_t) &\overset{(54)}{=} \mathcal{J}_{R,d}(\boldsymbol{x}_t)^\top \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t), r_t) + \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t)\right\|^2/R^2\right)\boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t) \\
&\overset{(57)}{=} \mathcal{J}_{R,d}(\boldsymbol{\rho}_{R,d}^{-1}(\boldsymbol{y}_t))^\top \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{y}_t, r_t) + \eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\boldsymbol{y}_t \\
&= \mathcal{G}(\boldsymbol{y}_t, \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{y}_t, r_t)),
\end{aligned} \tag{58}$$

where $\mathcal{G}(\boldsymbol{u}, \boldsymbol{v}) : \mathsf{int}\mathcal{B}^d(R) \times \mathbb{R}^d \to \mathbb{R}^d$ is a measurable mapping defined as

$$\mathcal{G}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \mathcal{J}_{R,d}(\boldsymbol{\rho}_{R,d}^{-1}(\boldsymbol{u}))^\top \boldsymbol{v} + \eta\left(2 - \|\boldsymbol{u}\|^2/R^2\right)\boldsymbol{u}.$$

Thus, we can write

$$\boldsymbol{y}_t \overset{(56),(57),(58)}{=} \boldsymbol{\rho}_{R,d} \circ \mathsf{A}_t\left(s, \mathcal{G}(\boldsymbol{y}_{t-1}, \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{y}_{t-1}, r_{t-1})), \ldots, \mathcal{G}(\boldsymbol{y}_1, \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{y}_1, r_1))\right), \forall t \in \mathbb{N}.$$

By a simple induction, there exists a sequence of measurable mappings $\mathsf{A}_t^{\boldsymbol{y}}, \forall t \in \mathbb{N}$ such that

$$\boldsymbol{y}_t = \mathsf{A}_t^{\boldsymbol{y}}(s, \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{y}_{t-1}, r_{t-1}), \ldots, \bar{\boldsymbol{g}}_{T,U}(\boldsymbol{y}_1, r_1)), \forall t \in \mathbb{N}.$$

The above reformulation implies $\boldsymbol{y}_t, \forall t \in \mathbb{N}$ can be viewed as a sequence produced by a randomized algorithm $\mathsf{A}^{\boldsymbol{y}} \in \mathsf{A}_{\mathsf{rand}}$ interacting with stochastic first-order oracle $\bar{\boldsymbol{g}}_{T,U}$. Note that $d \geq T + 32R^2 \log \frac{2T^2}{q\iota}$ and $\|\boldsymbol{y}_t\| \overset{(57)}{\leq} R$, we hence have by Lemma 4,

$$\Pr\left[\mathsf{prog}_{1/4}\left(U^\top \boldsymbol{y}_t\right) < T, \forall t \leq \frac{T - \log(2/\iota)}{2q}\right] \geq 1 - \iota. \tag{59}$$

Now we recall

$$\begin{aligned}
\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t) &\overset{(52)}{=} \mathcal{J}_{R,d}(\boldsymbol{x}_t)^\top \nabla \bar{F}_{T,U}(\boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t)) + \eta\left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t)\right\|^2/R^2\right)\boldsymbol{\rho}_{R,d}(\boldsymbol{x}_t) \\
&\overset{(57)}{=} \mathcal{J}_{R,d}(\boldsymbol{x}_t)^\top \nabla \bar{F}_{T,U}(\boldsymbol{y}_t) + \eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\boldsymbol{y}_t.
\end{aligned} \tag{60}$$

Moreover, by the first requirement on F_T, there is

$$\left\|\nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\| \overset{(48)}{=} \left\|U \nabla F_T(U^\top \boldsymbol{y}_t)\right\| \overset{U^\top U = I_T}{=} \left\|\nabla F_T(U^\top \boldsymbol{y}_t)\right\| \leq \gamma\sqrt{T}. \tag{61}$$

We fix $t \leq \frac{T - \log(2/\iota)}{2q}$ and consider the following two cases:

Case 1. $\|\boldsymbol{x}_t\| \geq \frac{R}{2}$. In this case, we first have

$$\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\| \overset{(60)}{\geq} \eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\|\boldsymbol{y}_t\| - \left\|\mathcal{J}_{R,d}(\boldsymbol{x}_t)^\top \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\|.$$

Note that $\frac{\|\boldsymbol{y}_t\|}{R} \overset{(57)}{=} \frac{\|\boldsymbol{x}_t\|/R}{\sqrt{1 + \|\boldsymbol{x}_t\|^2/R^2}} \in \left[\frac{1}{\sqrt{5}}, 1\right]$ when $\|\boldsymbol{x}_t\| \geq \frac{R}{2}$, implying

$$\eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\|\boldsymbol{y}_t\| \geq \eta R \min_{\frac{1}{\sqrt{5}} \leq a \leq 1} (2 - a^2)a = \frac{9\eta R}{5\sqrt{5}}.$$

In addition, there is

$$\left\|\mathcal{J}_{R,d}(\boldsymbol{x}_t)^\top \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\| \le \|\mathcal{J}_{R,d}(\boldsymbol{x}_t)\|_{\mathsf{op}} \left\|\nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\| \overset{\text{Lemma 5}}{=} \frac{\left\|\nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\|}{\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}}$$
$$\overset{\|\boldsymbol{x}_t\|\ge \frac{R}{2}}{\le} \frac{2\left\|\nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\|}{\sqrt{5}} \overset{(61)}{\le} \frac{2\gamma\sqrt{T}}{\sqrt{5}}.$$

As such, by our choices of η and R,

$$\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\| \ge \frac{9\eta R}{5\sqrt{5}} - \frac{2\gamma\sqrt{T}}{\sqrt{5}} \ge \frac{1}{2}.$$

Case 2. $\|\boldsymbol{x}_t\| < \frac{R}{2}$. In this case, we introduce

$$j_t \triangleq \mathsf{prog}_1\left(U^\top \boldsymbol{y}_t\right) + 1 \le \mathsf{prog}_{1/4}\left(U^\top \boldsymbol{y}_t\right) + 1 \overset{(59)}{\le} T.$$

Let $\boldsymbol{u}_{j_t} \in \mathbb{R}^d$ denotes the j_t-th column of U. By the definiton of j_t, there is

$$|\langle \boldsymbol{u}_{j_t}, \boldsymbol{y}_t\rangle| = \left|\left(U^\top \boldsymbol{y}_t\right)[j_t]\right| < 1. \tag{62}$$

In addition, we have

$$\left|\left\langle \boldsymbol{u}_{j_t}, \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\rangle\right| = \left|\left(U^\top \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right)[j_t]\right| \overset{(48), U^\top U = I_T}{=} \left|\nabla F_T(U^\top \boldsymbol{y}_t)[j_t]\right| > 1, \tag{63}$$

where the last step is by the second requirement on F_T. Now we compute

$$\begin{aligned}
&\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\| \overset{\|\boldsymbol{u}_{j_t}\|=1}{\ge} \left|\left\langle \boldsymbol{u}_{j_t}, \nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\rangle\right| \\
&\overset{(60)}{=} \left|\left\langle \boldsymbol{u}_{j_t}, \mathcal{J}_{R,d}(\boldsymbol{x}_t)^\top \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\rangle + \eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\langle \boldsymbol{u}_{j_t}, \boldsymbol{y}_t\rangle\right| \\
&\overset{\text{Lemma 5}}{=} \left|\frac{\left\langle \boldsymbol{u}_{j_t}, \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\rangle}{\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} - \frac{\langle \boldsymbol{u}_{j_t}, \boldsymbol{y}_t\rangle\left\langle \boldsymbol{y}_t, \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\rangle}{R^2\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} + \eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\langle \boldsymbol{u}_{j_t}, \boldsymbol{y}_t\rangle\right| \\
&\ge \frac{\left|\left\langle \boldsymbol{u}_{j_t}, \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\rangle\right|}{\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} - |\langle \boldsymbol{u}_{j_t}, \boldsymbol{y}_t\rangle|\left[\frac{\left|\left\langle \boldsymbol{y}_t, \nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\rangle\right|}{R^2\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} + \eta\left(2 - \|\boldsymbol{y}_t\|^2/R^2\right)\right] \\
&\overset{(a)}{\ge} \frac{1}{\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} - \frac{\|\boldsymbol{y}_t\|\left\|\nabla \bar{F}_{T,U}(\boldsymbol{y}_t)\right\|}{R^2\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} - 2\eta \overset{(b)}{\ge} \frac{2}{\sqrt{5}} - \frac{2\gamma\sqrt{T}}{5R} - 2\eta \overset{(c)}{=} \frac{1}{2},
\end{aligned}$$

where (a) is due to (62), (63), Cauchy-Schwarz inequality and $\eta \ge 0$, (b) is by $\frac{1}{\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} > \frac{2}{\sqrt{5}}$ and $\frac{\|\boldsymbol{y}_t\|}{\sqrt{1+\|\boldsymbol{x}_t\|^2/R^2}} \overset{(57)}{=} \frac{\|\boldsymbol{x}_t\|}{1+\|\boldsymbol{x}_t\|^2/R^2} < \frac{2R}{5}$ when $\|\boldsymbol{x}_t\| < \frac{R}{2}$ and (61), and (c) holds under our choices of η and R.

We combine two cases to conclude

$$\Pr\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\| \ge \frac{1}{2}, \forall t \le \frac{T - \log(2/\iota)}{2q}\right] \ge 1 - \iota.$$

□

Lastly, we recall the following hard instance and its stochastic first-order oracle studied in [3].

Lemma 7 (Lemma 2 of [3]). *For any $T \in \mathbb{N}$, there exists a differentiable function $F_T : \mathbb{R}^T \to \mathbb{R}$ satisfying the following properties:*

1. $F_T(\mathbf{0}) = 0$ *and* $\inf_{\boldsymbol{x}\in\mathbb{R}^T} F_T(\boldsymbol{x}) \ge -fT$*, where* $f = 12$.
2. F_T *is* ℓ*-smooth, where* $\ell = 152$.
3. *For all* $\boldsymbol{x} \in \mathbb{R}^T$, $\|\nabla F_T(\boldsymbol{x})\|_\infty \le \gamma$*, where* $\gamma = 23$.
4. *For all* $\boldsymbol{x} \in \mathbb{R}^T$, $\mathsf{prog}_0(\nabla F_T(\boldsymbol{x})) \le \mathsf{prog}_{1/2}(\boldsymbol{x}) + 1$.
5. *For all* $\boldsymbol{x} \in \mathbb{R}^T$ *and* $i \triangleq \mathsf{prog}_{1/2}(\boldsymbol{x})$, $\nabla F_T(\boldsymbol{x}) = \nabla F_T(\boldsymbol{x}_{\le i+1})$ *and* $[\nabla F_T(\boldsymbol{x})]_{\le i} = [\nabla F_T(\boldsymbol{x}_{\le i})]_{\le i}$.
6. *For all* $\boldsymbol{x} \in \mathbb{R}^T$*, if* $\mathsf{prog}_1(\boldsymbol{x}) < T$*, then* $\|\nabla F_T(\boldsymbol{x})\| \ge |\nabla F_T(\boldsymbol{x})\,[\mathsf{prog}_1(\boldsymbol{x}) + 1]| > 1$.

Lemma 8 (Lemma 3 of [3]). *For any $T \in \mathbb{N}$ and F_T in Lemma 7, the following $\boldsymbol{g}_T : \mathbb{R}^T \times \mathcal{R} \to \mathbb{R}^T$, where $\mathcal{R} \triangleq \{0, 1\}$, is a stochastic first-order oracle of F_T:*

$$\boldsymbol{g}_T(\boldsymbol{x}, r)\,[i] \triangleq \begin{cases} \nabla F_T(\boldsymbol{x})\,[i] & i \ne \mathsf{prog}_{1/4}(\boldsymbol{x}) + 1 \\ \frac{r}{q}\nabla F_T(\boldsymbol{x})\,[i] & i = \mathsf{prog}_{1/4}(\boldsymbol{x}) + 1 \end{cases}, \forall i \in [T]\,,$$

where $r = \mathsf{Bernoulli}(q)$ for some $q \in (0, 1)$. Moreover, $\boldsymbol{g}_T$ is a probability-q zero-chain.

E.6.3 Analysis under Heavy-Tailed Noise

From now on, we need to diverge from [3] and [7], since both of which are under the finite variance case (i.e., $\mathsf{p} = 2$) instead of heavy-tailed noise (i.e., $\mathsf{p} \in (1, 2]$).

Lemma 9 (variation of Lemma 7 in [3] and Lemma 26 in [7]). *The instances $\widehat{F}_{T,U}$ and $\widehat{\boldsymbol{g}}_{T,U}$ (see (49) and (50)) constucted based on F_T in Lemma 7 and $\boldsymbol{g}_T$ in Lemma 8 under $\eta = \frac{1}{\sqrt{5}} - \frac{39}{140}$ and $R = 7\gamma\sqrt{T}$ for $\gamma = 23$ satisfy the following properties:*

1. $\widehat{F}_{T,U}(\mathbf{0}) - \inf_{\boldsymbol{x}\in\mathbb{R}^d} \widehat{F}_{T,U}(\boldsymbol{x}) \le \widehat{f}T$*, where* $\widehat{f} = 12$.
2. $\widehat{F}_{T,U}$ *is* $\widehat{\ell}$*-smooth, where* $\widehat{\ell} = 154$.
3. *For all* $\boldsymbol{x} \in \mathbb{R}^d$, $\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x})\right\| \le \widehat{\gamma}\sqrt{T}$*, where* $\widehat{\gamma} = 53$.
4. *For all* $\boldsymbol{x} \in \mathbb{R}^d$ *and* $\mathsf{p} \in (1, 2]$, $\mathbb{E}_r\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}) - \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r)\right\|^{\mathsf{p}}\right] \le \frac{(2\gamma)^{\mathsf{p}}\left(1-q^{\mathsf{p}-1}\right)}{(\mathsf{p}-1)q^{\mathsf{p}-1}}$.

Proof. First, we know

$$\begin{aligned} \widehat{F}_{T,U}(\mathbf{0}) - \inf_{\boldsymbol{x}\in\mathbb{R}^d} \widehat{F}_{T,U}(\boldsymbol{x}) &\overset{(51)}{=} F_T(\mathbf{0}) - \inf_{\boldsymbol{x}\in\mathbb{R}^d} \left(F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta \boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right) \\ &\overset{\boldsymbol{x}^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}) \ge 0}{\le} F_T(\mathbf{0}) - \inf_{\boldsymbol{x}\in\mathbb{R}^T} F_T(\boldsymbol{x}) \overset{\text{Lemma 7}}{\le} fT = \widehat{f}T. \end{aligned}$$

Next, for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\nabla \widehat{F}_{T,U}(\boldsymbol{x}) \overset{(53)}{=} \mathcal{J}_{R,d}(\boldsymbol{x})^\top U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) + \eta\left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x}).$$

By Lemma 14 in [3] and Lemma 7, $\mathcal{J}_{R,d}(\boldsymbol{x})^\top U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}))$ is $\ell + \frac{3\gamma\sqrt{T}}{R} \overset{R=7\gamma\sqrt{T}}{=} \ell + \frac{3}{7}$-Lipschitz for $\ell = 152$. Moreover, we have

$$\begin{aligned}
&\left\| \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{x}) - \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{y})\right\|^2 / R^2\right) \boldsymbol{\rho}_{R,d}(\boldsymbol{y}) \right\| \\
\leq & 2 \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x}) - \boldsymbol{\rho}_{R,d}(\boldsymbol{y})\right\| + \frac{\left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{y})\right\|^2}{R^2} \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{y}) - \boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\| \\
& + \frac{\left| \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{y})\right\|^2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 \right|}{R^2} \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\| \\
\overset{(a)}{\leq} & 2 \left\|\boldsymbol{x} - \boldsymbol{y}\right\| + \left\|\boldsymbol{y} - \boldsymbol{x}\right\| + 2 \left\|\boldsymbol{y} - \boldsymbol{x}\right\| = 5 \left\|\boldsymbol{x} - \boldsymbol{y}\right\|,
\end{aligned}$$

where (a) is by $\left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x}) - \boldsymbol{\rho}_{R,d}(\boldsymbol{y})\right\| \overset{\text{Lemma 5}}{\leq} \left\|\boldsymbol{x} - \boldsymbol{y}\right\|$ and $\left\|\boldsymbol{\rho}_{R,d}(\cdot)\right\| \leq R$. So $\widehat{F}_{T,U}(\boldsymbol{x})$ is $\ell + \frac{3}{7} + 5\eta \overset{\eta = \frac{1}{\sqrt{5}} - \frac{39}{140}}{=} \frac{4229}{28} + \sqrt{5} \leq 154 = \widehat{\ell}$-smooth.

Moreover, we observe that

$$\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x})\right\| \overset{(53)}{\leq} \left\|\mathcal{J}_{R,d}(\boldsymbol{x})\right\|_{\mathsf{op}} \left\|U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}))\right\| + \eta \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|.$$

Note that for $\gamma = 23$,

$$\begin{aligned}
\left\|\mathcal{J}_{R,d}(\boldsymbol{x})\right\|_{\mathsf{op}} \left\|U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}))\right\| &\overset{\text{Lemma 5}}{\leq} \left\|U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}))\right\| \\
&\overset{U^\top U = I_T}{=} \left\|\nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}))\right\| \overset{\text{Lemma 7}}{\leq} \gamma\sqrt{T},
\end{aligned}$$

and

$$\begin{aligned}
\eta \left(2 - \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\|^2 / R^2\right) \left\|\boldsymbol{\rho}_{R,d}(\boldsymbol{x})\right\| &\overset{\left\|\boldsymbol{\rho}_{R,d}(\cdot)\right\| \leq R}{\leq} \eta R \max_{0 \leq a \leq 1} (2 - a^2) a = \eta R \cdot \frac{4}{3}\sqrt{\frac{2}{3}} \\
&= \left(\frac{1}{\sqrt{5}} - \frac{39}{140}\right) \cdot 7 \cdot \frac{4}{3}\sqrt{\frac{2}{3}} \cdot \gamma\sqrt{T} \leq 1.3\gamma\sqrt{T}.
\end{aligned}$$

We hence have $\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x})\right\| \leq 2.3\gamma\sqrt{T} \leq 53\sqrt{T} = \widehat{\gamma}\sqrt{T}$.

Lastly, still for $\gamma = 23$, we compute

$$\begin{aligned}
\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}) - \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r)\right\| &\overset{(53),(55)}{=} \left\|\mathcal{J}_{R,d}(\boldsymbol{x})^\top U \nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) - \mathcal{J}_{R,d}(\boldsymbol{x})^\top U \boldsymbol{g}_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r)\right\| \\
&\leq \left\|\mathcal{J}_{R,d}(\boldsymbol{x})\right\|_{\mathsf{op}} \left\|U \left(\nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) - \boldsymbol{g}_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r)\right)\right\| \\
&\overset{\text{Lemma 5}}{\leq} \left\|U \left(\nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) - \boldsymbol{g}_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r)\right)\right\| \\
&\overset{U^\top U = I_T}{=} \left\|\nabla F_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x})) - \boldsymbol{g}_T(U^\top \boldsymbol{\rho}_{R,d}(\boldsymbol{x}), r)\right\| \\
&\overset{\text{Lemma 8}}{=} \left|1 - \frac{r}{q}\right| \left|\nabla F_T(\boldsymbol{x}) \left[\mathsf{prog}_{1/4}(\boldsymbol{x}) + 1\right]\right| \overset{\text{Lemma 7}}{\leq} \gamma \left|1 - \frac{r}{q}\right|,
\end{aligned}$$

which implies

$$\begin{aligned}
\mathbb{E}_r\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}) - \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r)\right\|^{\mathsf{p}}\right] &\leq \gamma^{\mathsf{p}} \left(1 - q + \frac{(1-q)^{\mathsf{p}}}{q^{\mathsf{p}-1}}\right) = \gamma^{\mathsf{p}} (1-q) \left(\frac{q^{\mathsf{p}-1} + (1-q)^{\mathsf{p}-1}}{q^{\mathsf{p}-1}}\right) \\
&\overset{\text{Fact 3}}{\leq} \frac{\gamma^{\mathsf{p}} 2^{2-\mathsf{p}} \left(1 - q^{\mathsf{p}-1}\right)}{(\mathsf{p}-1) q^{\mathsf{p}-1}} \overset{\mathsf{p} > 1}{\leq} \frac{(2\gamma)^{\mathsf{p}} \left(1 - q^{\mathsf{p}-1}\right)}{(\mathsf{p}-1) q^{\mathsf{p}-1}}.
\end{aligned}$$

□

Next, inspired by [7], we first prove a lower bound for heavy-tailed smooth nonconvex optimization, as presented in the following Theorem 11.

Theorem 11. *For any $\Delta > 0$, $H > 0$, $\mathsf{p} \in (1,2]$, $\sigma \geq 0$, $0 < \epsilon \leq \sqrt{\frac{\Delta H}{96 \cdot 12 \cdot 154}}$, let d be in the order of $\frac{\Delta H}{\epsilon^2} \log\left(\frac{\Delta H}{\epsilon^2}\left(1 + \left(\frac{\sigma}{\epsilon}\right)^{\frac{\mathsf{p}}{2(\mathsf{p}-1)}}\right)\right)$ (see proof for the precise definition), there exists a distribution over functions $F : \mathbb{R}^d \to \mathbb{R}$ and stochastic first-order oracles $\boldsymbol{g}$ such that with probability 1, $F(\mathbf{0}) - F_\star \leq \Delta$, F is H-smooth and $\frac{3}{2}\sqrt{H\Delta}$-Lipschitz, and $\boldsymbol{g}$ has a finite p-th centered moment σ^{p}. Moreover, for any randomized $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$ employed to optimize a randomly selected F interacting with $\boldsymbol{g}$, to output a point $\boldsymbol{x}$ such that $\mathbb{E}\left[\|\nabla F(\boldsymbol{x})\|\right] \leq \epsilon$, the number of queries of $\boldsymbol{g}$ by A satisfies $\gtrsim \Delta H \epsilon^{-2} + \Delta H \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}} \epsilon^{-\frac{3\mathsf{p}-2}{\mathsf{p}-1}}$.*

Remark 10. The reader familiar with the literature may find that a similar lower bound (in fact, exactly the same order) has been established by [23, 49] before, and hence may wonder about the difference. Here, we note that the lower bounds in [23, 49] are shown for a special algorithmic class known as zero-respecting algorithms[4]. However, our Theorem 11 is proved for a broader family, i.e., randomized algorithms. This fact is important because Algorithm 4 is a randomized algorithm but not a zero-respecting algorithm.

Proof. For $T \in \mathbb{N}$ and $q \in (0,1)$ being determined later, let $\iota = 1/2$, $\eta = \frac{1}{\sqrt{5}} - \frac{39}{140}$, $R = 7\gamma\sqrt{T}$ where $\gamma = 23$ and $d = \left\lceil T + 32R^2 \log \frac{2T^2}{q\iota} \right\rceil = \left\lceil T + 32R^2 \log \frac{4T^2}{q} \right\rceil = \Theta(T \log \frac{T^2}{q})$, we construct $\widehat{F}_{T,U} : \mathbb{R}^d \to \mathbb{R}$ and $\widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r) : \mathbb{R}^d \times \mathcal{R} \to \mathbb{R}^d$ based on F_T in Lemma 7 and $\boldsymbol{g}_T$ in Lemma 8.

By Lemma 6 (note that F_T satisfies the requirements of Lemma 6 due to Lemma 7), for any $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$ employed to optimize $\widehat{F}_{T,U}$ interacting with the stochastic first-order oracle $\widehat{\boldsymbol{g}}_{T,U}$, where U is drawn from $\mathsf{Ortho}(d, T)$ uniformly, we have

$$\Pr\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t)\right\| \geq \frac{1}{2}, \forall t \leq \frac{T - \log 4}{2q}\right] \geq \frac{1}{2}. \tag{64}$$

By Lemma 9, $\widehat{F}_{T,U}(\mathbf{0}) - \inf_{\boldsymbol{x} \in \mathbb{R}^d} \widehat{F}_{T,U}(\boldsymbol{x}) \leq \widehat{f}T$ for $\widehat{f} = 12$, $\widehat{F}_{T,U}$ is $\widehat{\ell}$-smooth for $\widehat{\ell} = 154$ and $\widehat{\gamma}\sqrt{T}$-Lipschitz for $\widehat{\gamma} = 53$, $\mathbb{E}_r\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}) - \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}, r)\right\|^{\mathsf{p}}\right] \leq \frac{(2\gamma)^{\mathsf{p}}\left(1 - q^{\mathsf{p}-1}\right)}{(\mathsf{p}-1)q^{\mathsf{p}-1}}, \forall \boldsymbol{x} \in \mathbb{R}^d$.

Now we set

$$\lambda \triangleq \frac{4\widehat{\ell}\epsilon}{H}, \quad T \triangleq \left\lfloor \frac{\widehat{\ell}\Delta}{\widehat{f}H\lambda^2} \right\rfloor, \quad q \triangleq \frac{1}{\left[1 + (\mathsf{p}-1)\left(\frac{\widehat{\ell}\sigma}{2\gamma H\lambda}\right)^{\mathsf{p}}\right]^{\frac{1}{\mathsf{p}-1}}}. \tag{65}$$

For any $U \in \mathsf{Ortho}(d, T)$, we introduce

$$F_U(\boldsymbol{x}) \triangleq \frac{H\lambda^2}{\widehat{\ell}} \widehat{F}_{T,U}(\boldsymbol{x}/\lambda) \quad \text{and} \quad \boldsymbol{g}_U(\boldsymbol{x}) = \frac{H\lambda}{\widehat{\ell}} \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}/\lambda, r).$$

We observe that

$$F_U(\mathbf{0}) - \inf_{\boldsymbol{x} \in \mathbb{R}^d} F_U(\boldsymbol{x}) = \frac{H\lambda^2}{\widehat{\ell}}\left(\widehat{F}_{T,U}(\mathbf{0}) - \inf_{\boldsymbol{x} \in \mathbb{R}^d} \widehat{F}_{T,U}(\boldsymbol{x}/\lambda)\right) \leq \frac{H\lambda^2}{\widehat{\ell}} \widehat{f}T \overset{(65)}{\leq} \Delta.$$

Moreover, $\nabla F_U(\boldsymbol{x}) = \frac{H\lambda}{\widehat{\ell}} \nabla \widehat{F}_{T,U}(\boldsymbol{x}/\lambda)$, which implies F_U is H-smooth and $\frac{H\lambda\widehat{\gamma}\sqrt{T}}{\widehat{\ell}} \overset{(65)}{\leq} \frac{3\sqrt{H\Delta}}{2}$-Lipschitz. In addition, there is

$$\begin{aligned}
\mathbb{E}_r\left[\|\nabla F_U(\boldsymbol{x}) - \boldsymbol{g}_U(\boldsymbol{x})\|^{\mathsf{p}}\right] &= \left(\frac{H\lambda}{\widehat{\ell}}\right)^{\mathsf{p}} \mathbb{E}_r\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}/\lambda) - \widehat{\boldsymbol{g}}_{T,U}(\boldsymbol{x}/\lambda, r)\right\|^{\mathsf{p}}\right] \\
&\leq \left(\frac{2\gamma H\lambda}{\widehat{\ell}}\right)^{\mathsf{p}} \frac{1 - q^{\mathsf{p}-1}}{(\mathsf{p}-1)q^{\mathsf{p}-1}} \overset{(65)}{=} \sigma^{\mathsf{p}}.
\end{aligned}$$

[4]A first-order algorithm is called zero-respecting if it satisfies $\boldsymbol{x}_t \in \cup_{s<t} \text{support}(\boldsymbol{g}_s), \forall t \in \mathbb{N}$. For more details, see Definition 1 of [3].

For any $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$ used to optimize F_U with $\mathsf{StocGrad}_U$ when U is drawn from $\mathsf{Ortho}(d, T)$ uniformly, we can view as $\boldsymbol{x}_t/\lambda$ as the output of another $\mathsf{A}^\lambda \in \mathsf{A}_{\mathsf{rand}}$ interacting with $\widehat{F}_{T,U}$ and $\widehat{\boldsymbol{g}}_{T,U}$ (a similar argument is used in the proof of Lemma 6). As such, by (64),

$$\Pr\left[\left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t/\lambda)\right\| \geq \frac{1}{2}, \forall t \leq \frac{T - \log 4}{2q}\right] \geq \frac{1}{2},$$

which implies with probability at least 1/2

$$\|\nabla F_U(\boldsymbol{x}_t)\| = \frac{H\lambda}{\widehat{\ell}} \left\|\nabla \widehat{F}_{T,U}(\boldsymbol{x}_t/\lambda)\right\| \geq \frac{H\lambda}{2\widehat{\ell}}, \forall t \leq \frac{T - \log 4}{2q}.$$

Therefore, we have

$$\mathbb{E}\left[\|\nabla F_U(\boldsymbol{x}_t)\|\right] \geq \frac{H\lambda}{2\widehat{\ell}} \Pr\left[\|\nabla F_U(\boldsymbol{x}_t)\| \geq \frac{H\lambda}{2\widehat{\ell}}\right] \geq \frac{H\lambda}{4\widehat{\ell}} \overset{(65)}{=} \epsilon, \forall t \leq \frac{T - \log 4}{2q}.$$

Lastly, we compute

$$\begin{aligned}
\frac{T - \log 4}{2q} &\overset{(65)}{=} \frac{1}{2}\left(\left\lfloor \frac{\Delta H}{16\widehat{f}\widehat{\ell}\epsilon^2} \right\rfloor - \log 4\right)\left[1 + (\mathsf{p}-1)\left(\frac{\widehat{\ell}\sigma}{2\gamma H\lambda}\right)^{\mathsf{p}}\right]^{\frac{1}{\mathsf{p}-1}} \\
&\overset{(a)}{\geq} \frac{1}{2}\left(\frac{\Delta H}{16\widehat{f}\widehat{\ell}\epsilon^2} - 3\right)\left[1 + (\mathsf{p}-1)\left(\frac{\widehat{\ell}\sigma}{2\gamma H\lambda}\right)^{\mathsf{p}}\right]^{\frac{1}{\mathsf{p}-1}} \\
&\overset{(b)}{\geq} \frac{\Delta H}{64\widehat{f}\widehat{\ell}\epsilon^2}\left(1 + (\mathsf{p}-1)^{\frac{1}{\mathsf{p}-1}}\left(\frac{\widehat{\ell}\sigma}{2\gamma H\lambda}\right)^{\frac{\mathsf{p}}{\mathsf{p}-1}}\right) \\
&\overset{(65)}{\gtrsim} \Delta H \epsilon^{-2} + \Delta H \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}} \epsilon^{-\frac{3\mathsf{p}-2}{\mathsf{p}-1}}.
\end{aligned}$$

where the (a) is by $\lfloor\cdot\rfloor \geq \cdot - 1$ and $\log 4 \leq 2$, and (b) holds due to the condition $\epsilon^2 \leq \frac{\Delta H}{96 \cdot 12 \cdot 154} = \frac{\Delta H}{96\widehat{f}\widehat{\ell}}$, implying $\frac{\Delta H}{16\widehat{f}\widehat{\ell}\epsilon^2} - 3 \geq \frac{\Delta H}{32\widehat{f}\widehat{\ell}\epsilon^2}$, and $(1+x)^{\frac{1}{\mathsf{p}-1}} \geq 1 + x^{\frac{1}{\mathsf{p}-1}}$ for $x \geq 0$ when $\mathsf{p} \in (1, 2]$. So to achieve $\mathbb{E}\left[\|\nabla F_U(\boldsymbol{x}_t)\|\right] < \epsilon$, it requires at least $\Delta H \epsilon^{-2} + \Delta H \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}} \epsilon^{-\frac{3\mathsf{p}-2}{\mathsf{p}-1}}$ many iterations. In particular, we note that $d = \Theta(T \log \frac{T^2}{q}) = \Theta\left(\frac{\Delta H}{\epsilon^2} \log\left(\frac{\Delta H}{\epsilon^2}\left(1 + \left(\frac{\sigma}{\epsilon}\right)^{\frac{\mathsf{p}}{2(\mathsf{p}-1)}}\right)\right)\right)$. □

Equipped with Theorem 11, we can show the lower bound on the distributional complexity [26] for nonsmooth nonconvex optimization with heavy tails in the following Theorem 12.

Theorem 12. *For any given $\Delta > 0$, $G > 0$, $\mathsf{p} \in (1, 2]$, $\sigma \geq 0$, $\delta > 0$ and $0 < \epsilon \leq \frac{\Delta}{96 \cdot 12 \cdot 154\delta} \wedge \frac{4G^2\delta}{9\Delta}$, let d be in the order of $\frac{\Delta}{\delta\epsilon} \log\left(\frac{\Delta}{\delta\epsilon}\left(1 + \left(\frac{\sigma}{\epsilon}\right)^{\frac{\mathsf{p}}{2(\mathsf{p}-1)}}\right)\right)$, there exists a distribution over functions $F : \mathbb{R}^d \to \mathbb{R}$, and stochastic first-order oracles $\boldsymbol{g}$ such that with probability 1, $F(\mathbf{0}) - F_\star \leq \Delta$, F is G-Lipschitz and $\boldsymbol{g}$ has a finite p-th centered moment σ^{p}. Moreover, for any randomized $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$ employed to optimize a randomly selected F interacting with $\boldsymbol{g}$, to output a point $\boldsymbol{x}$ such that $\mathbb{E}\left[\|\nabla F(\boldsymbol{x})\|_\delta\right] \leq \epsilon$, the number of queries of $\boldsymbol{g}$ by A satisfies $\gtrsim \Delta\delta^{-1}\epsilon^{-1} + \Delta\sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$.*

Proof. Let $H \triangleq \epsilon/\delta$ in the following. Note that $\epsilon \leq \sqrt{\frac{\Delta H}{96 \cdot 12 \cdot 154}}$ due to our requirement on ϵ. So we can consider the same distribution on F and $\boldsymbol{g}$ as in Theorem 11. Importantly, F is H-smooth and $\frac{3}{2}\sqrt{H\Delta} = \frac{3}{2}\sqrt{\frac{\epsilon\Delta}{\delta}} \leq G$-Lipschitz (again due to the condition on ϵ) with probability 1.

If $\mathsf{A} \in \mathsf{A}_{\mathsf{rand}}$ finds a point $\boldsymbol{x}$ such that $\mathbb{E}\left[\|\nabla F(\boldsymbol{x})\|_\delta\right] \leq \epsilon$ (i.e., a (δ, ϵ)-stationary point of F), then by Proposition 14 of [7] (or $\nu = 1$ in Lemma 12 given later), it also satisfies $\mathbb{E}\left[\|\nabla F(\boldsymbol{x})\|\right] \leq \epsilon + H\delta = 2\epsilon$. Therefore, Theorem 11 implies the number of queries of $\boldsymbol{g}$ by A is at least $\gtrsim \Delta H \epsilon^{-2} + \Delta H \sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}} \epsilon^{-\frac{3\mathsf{p}-2}{\mathsf{p}-1}} = \Delta\delta^{-1}\epsilon^{-1} + \Delta\sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}}\delta^{-1}\epsilon^{-\frac{2\mathsf{p}-1}{\mathsf{p}-1}}$. □

Finally, we are able to prove Theorem 6.

Proof of Theorem 6. We recall the fact that lower bounds on the distributional complexity imply lower bounds on the minimax complexity [26]. Thus, Theorem 6 can be directly concluded from Theorem 12. ☐

F More Details about Further Extensions

This section contains more details of Section 5.

F.1 Full Results for Smooth ℓ_t

We present the full version of Theorem 7 and Corollary 4.

Theorem 13 (full version of Theorem 7)**.** *Under Assumption 1 (with replacing the third point by Condition 1) and let* $S_T(\boldsymbol{x}) \triangleq \left(H\sum_{t=1}^{T}\ell_t(\boldsymbol{x}) - \ell_t^{\inf}\right) \wedge \left(H\sum_{t=1}^{T}\ell_t(\boldsymbol{x}) - \ell_t^\star + \sum_{t=1}^{T}\|\nabla\ell_t(\boldsymbol{x}_t^\star)\|^2\right)$ *where* $\ell_t^{\inf} \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^d}\ell_t(\boldsymbol{x})$, $\boldsymbol{x}_t^\star \triangleq \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}}\ell_t(\boldsymbol{x})$, *and* $\ell_t^\star \triangleq \ell_t(\boldsymbol{x}_t^\star)$*:*

- *taking* $\eta_t = \frac{1}{4H} \wedge \gamma D \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ *for any* $\gamma > 0$ *in* $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$*, we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim HD^2 + D\left(\frac{1}{\gamma} + \gamma S_T(\boldsymbol{x})\right) + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x}\in\mathcal{X}.$$

- *taking* $\eta = D/\sqrt{2}$ *in* $\mathsf{A} = \mathsf{AdaGrad}$*, we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim HD^2 + D\sqrt{S_T(\boldsymbol{x})} + \sigma D T^{1/\mathsf{p}}, \forall \boldsymbol{x}\in\mathcal{X}.$$

Remark 11. First, $\boldsymbol{x}_t^\star$ must exist since $\mathcal{X}$ is a nonempty compact convex set and ℓ_t is closed convex. Next, when $\ell_t \geq 0$ on $\mathbb{R}^d$, we have $S_T(\boldsymbol{x}) \leq H\sum_{t=1}^{T}\ell_t(\boldsymbol{x}) - \ell_t^{\inf} \leq H\sum_{t=1}^{T}\ell_t(\boldsymbol{x})$. This fact, together with the replacement of γ by $\gamma/\sqrt{H}$ for OGD and DA, recovers Theorem 7.

Proof. Under Condition 1, the following famous inequality holds (regardless of the convexity of ℓ_t)

$$\|\nabla\ell_t(\boldsymbol{x})\|^2 \leq 2H(\ell_t(\boldsymbol{x}) - \ell_t^{\inf}), \forall \boldsymbol{x}\in\mathbb{R}^d. \tag{66}$$

For convex ℓ_t, another similar inequality is also true

$$\|\nabla\ell_t(\boldsymbol{x}) - \nabla\ell_t(\boldsymbol{x}_t^\star)\|^2 \leq 2H(\ell_t(\boldsymbol{x}) - \ell_t^\star), \forall \boldsymbol{x}\in\mathcal{X},$$

which implies

$$\|\nabla\ell_t(\boldsymbol{x})\|^2 \leq \left(\sqrt{2H(\ell_t(\boldsymbol{x}) - \ell_t^\star)} + \|\nabla\ell_t(\boldsymbol{x}_t^\star)\|\right)^2 \leq 3H(\ell_t(\boldsymbol{x}) - \ell_t^\star) + 3\|\nabla\ell_t(\boldsymbol{x}_t^\star)\|^2, \forall \boldsymbol{x}\in\mathcal{X}. \tag{67}$$

By (66) and (67), there is

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla\ell_t(\boldsymbol{x}_t)\|^2 &\leq 3\sum_{t=1}^{T}\left[H(\ell_t(\boldsymbol{x}_t) - \ell_t^{\inf}) \wedge \left(H(\ell_t(\boldsymbol{x}_t) - \ell_t^\star) + \|\nabla\ell_t(\boldsymbol{x}_t^\star)\|^2\right)\right] \\ &= 3H\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}) + 3\sum_{t=1}^{T}\left[H(\ell_t(\boldsymbol{x}) - \ell_t^{\inf}) \wedge \left(H(\ell_t(\boldsymbol{x}) - \ell_t^\star) + \|\nabla\ell_t(\boldsymbol{x}_t^\star)\|^2\right)\right] \\ &\leq 3H\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}) + 3S_T(\boldsymbol{x}).\end{aligned} \tag{68}$$

With (68) on hand, we can extend the regret bound of $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}, \mathsf{AdaGrad}\}$ to smooth ℓ_t.

For OGD, note that (13) still holds up to the change of G by $\|\nabla \ell_t(\boldsymbol{x}_t)\|$, meaning that for any $\boldsymbol{x} \in \mathcal{X}$,

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_t \|\nabla \ell_t(\boldsymbol{x}_t)\|^2 + \mathsf{C}(\mathsf{p}) \eta_t^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}. \tag{69}$$

For DA, note that (19) still holds up to the change of G by $\|\nabla \ell_t(\boldsymbol{x}_t)\|$, meaning that for any $\boldsymbol{x} \in \mathcal{X}$,

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_{t-1} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2 + \mathsf{C}(\mathsf{p}) \eta_{t-1}^{\mathsf{p}-1} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} D^{2-\mathsf{p}}, \tag{70}$$

where $\eta_0 = \eta_1$. Now we write $\eta_t = \eta \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ for $\eta = \frac{1}{4H} \wedge \gamma D$. For $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$, taking expectations on both sides of (69) or (69) implies

$$\begin{aligned}
\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] &\lesssim \frac{D^2}{\eta} + \sigma D T^{1/\mathsf{p}} + \eta \mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2\right] \\
&\overset{(68)}{\lesssim} \frac{D^2}{\eta} + \sigma D T^{1/\mathsf{p}} + 3\eta H \mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] + 3\eta S_T(\boldsymbol{x})
\end{aligned}$$

Since $\eta = \frac{1}{4H} \wedge \gamma D \Rightarrow 1 - 3\eta H \geq \frac{1}{4}$, $\eta \leq \gamma D$ and $\frac{1}{\eta} \leq 4H + \frac{1}{\gamma D}$, we hence conclude

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim \frac{D^2}{\eta(1-3\eta H)} + \frac{3\eta S_T(\boldsymbol{x})}{1-3\eta H} + \frac{\sigma D T^{1/\mathsf{p}}}{1-3\eta H} \lesssim H D^2 + D\left(\frac{1}{\gamma} + \gamma S_T(\boldsymbol{x})\right) + \sigma D T^{1/\mathsf{p}}.$$

For $\mathsf{A} = \mathsf{AdaGrad}$, note that (25) still holds, i.e., for any $\boldsymbol{x} \in \mathcal{X}$,

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \sqrt{2}\left(\frac{D^2}{2\eta} + \eta\right)\left[\sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2} + \left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right]. \tag{71}$$

We take expectations on both sides of (71), then apply Hölder's inequality to have $\mathbb{E}\left[\left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \leq \left(\sum_{t=1}^{T} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \leq \sigma T^{\frac{1}{\mathsf{p}}}$, and plug in $\eta = D/\sqrt{2}$ to obtain

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim D \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2}\right] + \sigma D T^{1/\mathsf{p}} \leq D\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \ell_t(\boldsymbol{x}_t)\|^2\right]} + \sigma D T^{1/\mathsf{p}},$$

where the last step is again by Hölder's inequality. Combine the above inequality with (68) to conclude

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim D\sqrt{H \mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] + S_T(\boldsymbol{x})} + \sigma D T^{1/\mathsf{p}} \\
\Rightarrow &\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim H D^2 + D\sqrt{S_T(\boldsymbol{x})} + \sigma D T^{1/\mathsf{p}}.
\end{aligned}$$

□

Corollary 7 (full version of Corollary 4)**.** *Under Assumption 1 (with replacing the third point by Condition 1) for $\ell_t(\boldsymbol{x}) = F(\boldsymbol{x})$ and let $\boldsymbol{x}^\star \triangleq \operatorname{argmin}_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x})$ and $\bar{\boldsymbol{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$, taking $\eta_t = \frac{1}{4H} \wedge \frac{D}{\|\nabla F(\boldsymbol{x}^\star)\|\sqrt{T}} \wedge \frac{D}{\sigma t^{1/\mathsf{p}}}$ in $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$ or $\eta = D/\sqrt{2}$ in $\mathsf{A} = \mathsf{AdaGrad}$, we have*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}^\star)\right] \leq \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}^\star)\right]}{T} \lesssim \frac{H D^2}{T} + \frac{\|\nabla F(\boldsymbol{x}^\star)\| D}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}.$$

Remark 12. When $\boldsymbol{x}^\star \in \mathcal{X}$ is also a global minimizer (i.e., $\boldsymbol{x}^\star \in \operatorname{arginf}_{\boldsymbol{x} \in \mathbb{R}^d} F(\boldsymbol{x})$), we have $\|\nabla F(\boldsymbol{x}^\star)\| = 0$ since F is differentiable. This fact recovers Theorem 7.

Proof. By convexity, $F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}^\star) \leq \frac{\sum_{t=1}^{T} F(\boldsymbol{x}_t) - F(\boldsymbol{x}^\star)}{T} = \frac{\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}^\star)}{T}$ is valid for any OCO algorithm A. We conclude from invoking Theorem 13 with $\gamma = \frac{1}{\|\nabla F(\boldsymbol{x}^\star)\|\sqrt{T}}$ for $\mathsf{A} \in \{\mathsf{OGD}, \mathsf{DA}\}$ and $\eta = D/\sqrt{2}$ for $\mathsf{A} = \mathsf{AdaGrad}$ and combining the fact $S_T(\boldsymbol{x}^\star) \leq H\sum_{t=1}^{T} F(\boldsymbol{x}^\star) - F(\boldsymbol{x}^\star) + \sum_{t=1}^{T} \|\nabla F(\boldsymbol{x}^\star)\|^2 = \|\nabla F(\boldsymbol{x}^\star)\|^2 T$ in this case. □

F.2 An Optimistic Algorithm under Heavy Tails

As discussed in Section 5, our goal is to handle broader cases with optimistic algorithms. To do so, we first introduce the following new Assumption 3:

Assumption 3. *We consider the following series of assumptions:*

- $\mathcal{X} \subset \mathbb{R}^d$ *is a nonempty closed convex set bounded by* D, *i.e.,* $\sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{y}\| \le D$.
- $\ell_t : \mathcal{X} \to \mathbb{R}$ *is closed convex for all* $t \in [T]$.
- ℓ_t *is* (G_t, H_t, ν)*-general nonsmooth on* $\mathcal{X}$, *i.e., there exists* $G_t \ge 0$, $H_t \ge 0$ *and* $\nu \in (0,1]$ *such that* $G_t + H_t > 0$ *and* $\|\nabla \ell_t(\boldsymbol{x}) - \nabla \ell_t(\boldsymbol{y})\| \le 2G_t + H_t \|\boldsymbol{x}-\boldsymbol{y}\|^{\nu}, \forall \boldsymbol{z} \in \mathcal{X}, \nabla \ell_t(\boldsymbol{z}) \in \partial \ell_t(\boldsymbol{z}), \boldsymbol{z} \in \{\boldsymbol{x},\boldsymbol{y}\}$, *for all* $t \in [T]$.
- *Given a point* $\boldsymbol{x}_t \in \mathcal{X}$ *at the t-th iteration, one can query* $\boldsymbol{g}_t \in \mathbb{R}^d$ *satisfying* $\nabla \ell_t(\boldsymbol{x}_t) \triangleq \mathbb{E}[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}] \in \partial \ell_t(\boldsymbol{x}_t)$ *and* $\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right] \le \sigma_t^{\mathsf{p}}$ *for some* $\mathsf{p} \in (1,2]$ *and* $\sigma_t \ge 0$, *where* $\mathcal{F}_t \triangleq \sigma(\boldsymbol{g}_1,\dots,\boldsymbol{g}_t)$ *denotes the natural filtration and* $\boldsymbol{\epsilon}_t \triangleq \boldsymbol{g}_t - \nabla \ell_t(\boldsymbol{x}_t)$ *is the stochastic noise.*

Assumption 3 generalizes Assumption 1 since the latter is a special case of the former, i.e., when $(G_t, H_t, \nu) = (G, 0, \nu)$ (for any $\nu \in (0,1]$) and $\sigma_t = \sigma$. When $G_t = 0$, this new assumption means that each ℓ_t is locally Hölder smooth with time-varying parameters (H_t, ν) on $\mathcal{X}$. If we further let $\nu = 1$, then each ℓ_t is standard smooth with a parameter H_t.

F.2.1 Optimistic AdaGrad

Algorithm 5 Optimistic AdaGrad (OAdaGrad)

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, initial hint $\boldsymbol{h}_1 \in \mathbb{R}^d$, stepsize $\eta > 0$ and $\gamma_t > 0$
for $t = 1$ **to** T **do**
  Query a hint $\boldsymbol{h}_{t+1}$
  $\eta_t = \eta V_t^{-1/2} \wedge \gamma_t$ where $V_t = \sum_{s=1}^{t} \|\boldsymbol{g}_s - \boldsymbol{h}_s\|^2$
  $\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} \langle \boldsymbol{g}_t + \boldsymbol{h}_{t+1} - \boldsymbol{h}_t, \boldsymbol{x}\rangle + \frac{\|\boldsymbol{x}-\boldsymbol{x}_t\|^2}{2\eta_t}$
end for

Following the famous idea of optimism, we consider an optimistic version of AdaGrad named OAdaGrad in Algorithm 5. To start with, we give the following Lemma 10. As a sanity check, one can take $\boldsymbol{h}_t = \mathbf{0}$ and $\gamma_t = +\infty$, then OAdaGrad reduces to the standard AdaGrad method, and Lemma 10 recovers the path-wise regret in (7). Therefore, the following result can be viewed as a further extension of AdaGrad, meaning that we can apply our idea described before (see the paragraph under (7) or the proof around (25)) to overcome heavy-tailed noise.

Lemma 10. *Under Assumption 3, taking any hint sequence* $\boldsymbol{h}_t \in \mathbb{R}^d$, $\eta = D/\sqrt{2}$, *and any nonincreasing stepsize* γ_t *in* OAdaGrad *(Algorithm 5), we have*

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \lesssim \langle \boldsymbol{h}_1, \boldsymbol{x}_1 - \boldsymbol{x}\rangle + \frac{D^2}{\gamma_T} + D\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t - \boldsymbol{h}_t\|^2} - \sum_{t=1}^{T-1} \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2}{4\gamma_t}, \forall \boldsymbol{x} \in \mathcal{X}.$$

Proof. Given $t \in [T]$, by the optimality condition of the update rule in Algorithm 5, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\left\langle \boldsymbol{g}_t + \boldsymbol{h}_{t+1} - \boldsymbol{h}_t + \frac{\boldsymbol{x}_{t+1} - \boldsymbol{x}_t}{\eta_t}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \right\rangle \le 0,$$

which implies

$$\begin{aligned}\langle \boldsymbol{g}_t + \boldsymbol{h}_{t+1} - \boldsymbol{h}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle &\le \frac{\langle \boldsymbol{x}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle}{\eta_t} \\ &= \frac{\|\boldsymbol{x}-\boldsymbol{x}_t\|^2 - \|\boldsymbol{x}-\boldsymbol{x}_{t+1}\|^2 - \|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{2\eta_t},\end{aligned}$$

Therefore, we know

$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle =& \langle \boldsymbol{g}_t + \boldsymbol{h}_{t+1} - \boldsymbol{h}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle + \langle \boldsymbol{g}_t - \boldsymbol{h}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle + \langle \boldsymbol{h}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle \\ \le& \frac{\|\boldsymbol{x}-\boldsymbol{x}_t\|^2 - \|\boldsymbol{x}-\boldsymbol{x}_{t+1}\|^2}{2\eta_t} + \langle \boldsymbol{g}_t - \boldsymbol{h}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1}\rangle - \frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{2\eta_t} \\ &+ \langle \boldsymbol{h}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle \\ \le& \frac{\|\boldsymbol{x}-\boldsymbol{x}_t\|^2 - \|\boldsymbol{x}-\boldsymbol{x}_{t+1}\|^2}{2\eta_t} + \eta_t \|\boldsymbol{g}_t - \boldsymbol{h}_t\|^2 - \frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\eta_t} \\ &+ \langle \boldsymbol{h}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle ,\end{aligned}$$

sum up which from $t = 1$ to T and drop the term $-\frac{\|\boldsymbol{x}_{T+1}-\boldsymbol{x}\|^2}{2\eta_T}$ to have

$$\begin{aligned}\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \le& \frac{\|\boldsymbol{x}-\boldsymbol{x}_1\|^2}{2\eta_1} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}\|^2}{2} + \sum_{t=1}^{T} \eta_t \|\boldsymbol{g}_t - \boldsymbol{h}_t\|^2 - \frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\eta_t} \\ &+ \langle \boldsymbol{h}_1, \boldsymbol{x}_1 - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{T+1}, \boldsymbol{x}_{T+1} - \boldsymbol{x}\rangle \\ \overset{(a)}{\le}& \frac{D^2}{2\eta_T} + \sum_{t=1}^{T} \eta_t \|\boldsymbol{g}_t - \boldsymbol{h}_t\|^2 - \frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\eta_t} + \langle \boldsymbol{h}_1, \boldsymbol{x}_1 - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{T+1}, \boldsymbol{x}_{T+1} - \boldsymbol{x}\rangle \\ \overset{(b)}{\le}& \frac{D^2}{2\gamma_T} + \frac{D^2\sqrt{V_T}}{2\eta} + \sum_{t=1}^{T} \frac{\eta \|\boldsymbol{g}_t - \boldsymbol{h}_t\|^2}{\sqrt{V_t}} - \frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\gamma_t} + \langle \boldsymbol{h}_1, \boldsymbol{x}_1 - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{T+1}, \boldsymbol{x}_{T+1} - \boldsymbol{x}\rangle \\ \overset{(c)}{\le}& \frac{D^2}{2\gamma_T} + \left(\frac{D^2}{2\eta} + 2\eta\right)\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t - \boldsymbol{h}_t\|^2} - \sum_{t=1}^{T} \frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\gamma_t} + \langle \boldsymbol{h}_1, \boldsymbol{x}_1 - \boldsymbol{x}\rangle - \langle \boldsymbol{h}_{T+1}, \boldsymbol{x}_{T+1} - \boldsymbol{x}\rangle ,\end{aligned}$$

where (a) is by $\|\boldsymbol{x}_t - \boldsymbol{x}\| \le D, \forall t \in [T]$ and $\eta_{t+1} \le \eta_t, \forall t \in [T-1]$ (since γ_t is assumed to be nonincreasing), (b) is due to $\eta_t = \frac{\eta}{\sqrt{V_t}} \wedge \gamma_t \Rightarrow \frac{1}{\eta_T} = \frac{\sqrt{V_T}}{\eta} \vee \frac{1}{\gamma_T} \le \frac{\sqrt{V_T}}{\eta} + \frac{1}{\gamma_T}$, and (c) follows a similar step as proving (24). Finally, we drop the term $-\frac{\|\boldsymbol{x}_{T+1}-\boldsymbol{x}_T\|^2}{4\gamma_T}$, use $\eta = D/\sqrt{2}$, and set $\boldsymbol{h}_{T+1} = \boldsymbol{0}$ to obtain the desired bound (this step is without loss of generality, otherwise, one can simply change the $\boldsymbol{x}_{T+1}$ used above to be $\operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} \langle \boldsymbol{g}_T - \boldsymbol{h}_T, \boldsymbol{x}\rangle + \frac{\|\boldsymbol{x}-\boldsymbol{x}_T\|^2}{2\eta_T}$). □

F.2.2 New Regret for OAdaGrad

Equipped with Lemma 10, we first prove the following Theorem 14, which establishes the regret of OAdaGrad under heavy tails.

Theorem 14. *Under Assumption 3, taking $\boldsymbol{h}_t = \boldsymbol{g}_{t-1}$ where $\boldsymbol{g}_0 \triangleq \boldsymbol{0}$, $\eta = D/\sqrt{2}$, and $\gamma_t = \frac{D^{1-\nu}}{\max_{s\in[t]} H_s}$ in* OAdaGrad *(Algorithm 5), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OAdaGrad}}(\boldsymbol{x})\right] \lesssim D^{1+\nu}\left(\sum_{t=1}^{T} H_t^{\frac{2-\nu}{1-\nu}}\right)^{\frac{1-\nu}{2-\nu}} + D\sqrt{A_T + \sum_{t=1}^{T} G_t^2} + D\left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},$$

where $A_T \triangleq \|\nabla \ell_1(\boldsymbol{x}_1)\| + \sum_{t=2}^{T} \sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla \ell_t(\boldsymbol{x}) - \nabla \ell_{t-1}(\boldsymbol{x})\|^2$ and $\ell_0 \triangleq 0$.

We briefly discuss Theorem 14 before proving it. First, the quantity A_T is standard and well-known in the literature as gradient variation [6]. Next, let us consider the case $H_t = 0$, then the bound degenerates to $D\sqrt{A_T + \sum_{t=1}^{T} G_t^2} + D\left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}}\right)^{1/\mathsf{p}}$, which further reduces to the optimal regret $GD\sqrt{T} + \sigma D T^{1/\mathsf{p}}$ under Assumption 1. In the case $G_t = 0$, i.e., locally Hölder smooth ℓ_t, it gives a regret $D^{1+\nu}\left(\sum_{t=1}^{T} H_t^{\frac{2-\nu}{1-\nu}}\right)^{\frac{1-\nu}{2-\nu}} +$

$D\sqrt{A_T} + D\left(\sum_{t=1}^{T}\sigma_t^{\mathsf{p}}\right)^{1/\mathsf{p}}$. In the special situation, deterministic OCO under standard smoothness (i.e., $\sigma_t = 0$, $H_t = H$, and $\nu = 1$), this matches the classical result $HD^2 + \sqrt{A_T}D$ [6]. But as one can see, our Theorem 14 is more general and, as far as we know, is the first bound containing gradient variation for heavy-tailed OCO.

Proof. By Lemma 10 with $\boldsymbol{h}_t = \boldsymbol{g}_{t-1}$, we have

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \lesssim \frac{D^2}{\gamma_T} + D\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}\right\|^2} - \sum_{t=1}^{T-1}\frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\gamma_t}. \tag{72}$$

Let $\boldsymbol{\epsilon}_0 \triangleq \mathbf{0}$, we can find

$$\begin{aligned}\sum_{t=1}^{T}\left\|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}\right\|^2 &= \sum_{t=1}^{T}\|\nabla\ell_t(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_t) + \nabla\ell_{t-1}(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_{t-1}) + \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{t-1}\|^2\\ &\lesssim \sum_{t=1}^{T}\|\nabla\ell_t(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_t)\|^2 + \sum_{t=1}^{T}\|\nabla\ell_{t-1}(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_{t-1})\|^2 + \sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^2\\ &\le A_T + \sum_{t=2}^{T}\|\nabla\ell_{t-1}(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_{t-1})\|^2 + \sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^2,\end{aligned} \tag{73}$$

where the last step is by the definition of A_T and $\ell_0 = 0$. Now by Assumption 3, $\|\nabla\ell_{t-1}(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_{t-1})\|^2 \lesssim G_{t-1}^2 + H_{t-1}^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^{2\nu}$, which implies

$$\sum_{t=2}^{T}\|\nabla\ell_{t-1}(\boldsymbol{x}_t) - \nabla\ell_{t-1}(\boldsymbol{x}_{t-1})\|^2 \lesssim \sum_{t=2}^{T}G_{t-1}^2 + H_{t-1}^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^{2\nu} = \sum_{t=1}^{T-1}G_t^2 + H_t^2\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^{2\nu}. \tag{74}$$

Combine (73) and (74) and use $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}, \forall a, b \ge 0$ to have

$$\begin{aligned}\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}\right\|^2} &\lesssim \sqrt{A_T + \sum_{t=1}^{T-1}G_t^2} + \sqrt{\sum_{t=1}^{T-1}H_t^2\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^{2\nu}} + \sqrt{\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^2}\\ &\le \sqrt{A_T + \sum_{t=1}^{T-1}G_t^2} + \sqrt{\sum_{t=1}^{T-1}H_t^2\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^{2\nu}} + \left(\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},\end{aligned} \tag{75}$$

where the last step is due to $\|\cdot\|_2 \le \|\cdot\|_{\mathsf{p}}$ for any $\mathsf{p} \in [1, 2]$.

Therefore, by (72) and (75), the following inequality holds

$$\begin{aligned}\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \lesssim &\frac{D^2}{\gamma_T} + D\sqrt{\sum_{t=1}^{T-1}H_t^2\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^{2\nu}} - \sum_{t=1}^{T-1}\frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{4\gamma_t}\\ &+ D\sqrt{A_T + \sum_{t=1}^{T-1}G_t^2} + D\left(\sum_{t=1}^{T}\|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}.\end{aligned} \tag{76}$$

We use Hölder's inequality to bound

$$\sum_{t=1}^{T-1}H_t^2\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^{2\nu} = \sum_{t=1}^{T-1}H_t^2(2\nu\gamma_t)^{\nu}\frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^{2\nu}}{(2\nu\gamma_t)^{\nu}} \le \left(\sum_{t=1}^{T-1}H_t^{\frac{2}{1-\nu}}(2\nu\gamma_t)^{\frac{\nu}{1-\nu}}\right)^{1-\nu}\left(\sum_{t=1}^{T-1}\frac{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2}{2\nu\gamma_t}\right)^{\nu},$$

which implies

$$D\sqrt{\sum_{t=1}^{T-1} H_t^2 \left\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\right\|^{2\nu}} \leq D\left(\sum_{t=1}^{T-1} H_t^{\frac{2}{1-\nu}}(2\nu\gamma_t)^{\frac{\nu}{1-\nu}}\right)^{\frac{1-\nu}{2}}\left(\sum_{t=1}^{T-1}\frac{\left\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\right\|^2}{2\nu\gamma_t}\right)^{\frac{\nu}{2}}$$

$$\leq \frac{D^{\frac{2}{2-\nu}}\left(\sum_{t=1}^{T-1} H_t^{\frac{2}{1-\nu}}(2\nu\gamma_t)^{\frac{\nu}{1-\nu}}\right)^{\frac{1-\nu}{2-\nu}}}{2/(2-\nu)} + \frac{\sum_{t=1}^{T-1}\frac{\left\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\right\|^2}{2\nu\gamma_t}}{2/\nu}$$

$$\lesssim D^{\frac{2}{2-\nu}}\left(\sum_{t=1}^{T-1} H_t^{\frac{2}{1-\nu}}\gamma_t^{\frac{\nu}{1-\nu}}\right)^{\frac{1-\nu}{2-\nu}} + \sum_{t=1}^{T-1}\frac{\left\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\right\|^2}{4\gamma_t}, \tag{77}$$

where the second step is due to Young's inequality. Plug (77) back into (76) to know

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t-\boldsymbol{x}\rangle \lesssim \frac{D^2}{\gamma_T} + D^{\frac{2}{2-\nu}}\left(\sum_{t=1}^{T-1} H_t^{\frac{2}{1-\nu}}\gamma_t^{\frac{\nu}{1-\nu}}\right)^{\frac{1-\nu}{2-\nu}} + D\sqrt{A_T+\sum_{t=1}^{T-1}G_t^2} + D\left(\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}.$$

Recall that $\gamma_t = \frac{D^{1-\nu}}{\max_{s\in[t]} H_s}$, we hence have

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t-\boldsymbol{x}\rangle \lesssim D^{1+\nu}\left(\sum_{t=1}^{T-1} H_t^{\frac{2-\nu}{1-\nu}}\right)^{\frac{1-\nu}{2-\nu}} + D\sqrt{A_T+\sum_{t=1}^{T-1}G_t^2} + D\left(\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}.$$

Finally, we conclude by taking expectations on both sides of the above inequality and using Hölder's inequality again to obtain

$$\mathbb{E}\left[\left(\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \leq \left(\mathbb{E}\left[\sum_{t=1}^{T}\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \leq D\left(\sum_{t=1}^{T}\sigma_t^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}.$$

□

An undesired point of the above Theorem 14 is requiring the knowledge of both ν and H_t to set the stepsize γ_t. In the following, we show that under a slightly variant nonsmooth notion, one can relax this requirement. To do so, we consider Condition 2 to substitute the third point in Assumption 3.

Condition 2. ℓ_t is $(G_t, H_t, \nu, \star)$-general nonsmooth on $\mathcal{X}$, i.e., there exists $G_t \geq 0$, $H_t \geq 0$ and $\nu \in (0,1]$ such that $G_t + H_t > 0$ and $\|\nabla\ell_t(\boldsymbol{x}) - \nabla\ell_t(\boldsymbol{x}_t^\star)\| \lesssim G_t + H_t^{\frac{1}{1+\nu}}(\ell_t(\boldsymbol{x}) - \ell_t^\star)^{\frac{\nu}{1+\nu}}, \forall \boldsymbol{x} \in \mathcal{X}, \nabla\ell_t(\boldsymbol{x}) \in \partial\ell_t(\boldsymbol{x})$ where $\boldsymbol{x}_t^\star \triangleq \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}}\ell_t(\boldsymbol{x})$ and $\ell_t^\star \triangleq \ell_t(\boldsymbol{x}_t^\star)$, for all $t \in [T]$.

To gain some intuition of the new Condition 2, we first let $H_t = 0$ to have $\|\nabla\ell_t(\boldsymbol{x}) - \nabla\ell_t(\boldsymbol{x}_t^\star)\| \lesssim G_t$, which is similar to $\|\nabla\ell_t(\boldsymbol{x}) - \nabla\ell_t(\boldsymbol{y})\| \lesssim G_t, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ in Assumption 3. Next, we consider the case $G_t = 0$, where Condition 2 states an inequality $\|\nabla\ell_t(\boldsymbol{x}) - \nabla\ell_t(\boldsymbol{x}_t^\star)\| \lesssim H_t^{\frac{1}{1+\nu}}(\ell_t(\boldsymbol{x}) - \ell_t^\star)^{\frac{\nu}{1+\nu}}$, which is known to hold if ℓ_t is (H_t, ν)-Hölder smooth on $\mathbb{R}^d$ and convex, which can be viewed as a generalization of Assumption 3 (though strictly speaking, we only require ℓ_t to be Hölder smooth on $\mathcal{X}$ in Assumption 3). For general G_t, H_t, and ν, whether Condition 2 is strictly general than the thrird point in Assumption 3 is unclear. Thus, we consider it as a separate condition here.

Now, we prove a new regret for OAdaGrad under Condition 2, which no longer needs to know H_t and ν.

Theorem 15. *Under Assumption 3 (with replacing the third point by Condition 2), taking $\boldsymbol{h}_t = \boldsymbol{g}_{t-1}$ where $\boldsymbol{g}_0 \triangleq \mathbf{0}$, $\eta = D/\sqrt{2}$, and $\gamma_t = +\infty$ in* OAdaGrad *(Algorithm 5), we have*

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{OAdaGrad}}(\boldsymbol{x})\right] \lesssim D^{1+\nu}\left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}}\right)^{\frac{1-\nu}{2}} + D\left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}}\right)^{\frac{1-\nu}{2(1+\nu)}}\left[C_T(\boldsymbol{x})\right]^{\frac{\nu}{1+\nu}} + D\sqrt{B_T+\sum_{t=1}^{T}G_t^2} + D\left(\sum_{t=1}^{T}\sigma_t^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},$$

where $B_T \triangleq \sum_{t=1}^{T}\left\|\nabla\ell_t(\boldsymbol{x}_t^\star) - \nabla\ell_{t-1}(\boldsymbol{x}_{t-1}^\star)\right\|^2$, $C_T(\boldsymbol{x}) \triangleq \sum_{t=1}^{T}\ell_t(\boldsymbol{x}) - \ell_t^\star$, and $\ell_0 \triangleq 0$.

Note that Theorem 15 and the previous Theorem 14 are not directly comparable due to different assumptions. But we can consider some special cases to better understand the difference. For example, let $H_t = 0$, Theorem 15 degenerates to $D\sqrt{B_T + \sum_{t=1}^T G_t^2} + D\left(\sum_{t=1}^T \sigma_t^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}$, similar to the regret $D\sqrt{A_T + \sum_{t=1}^T G_t^2} + D\left(\sum_{t=1}^T \sigma_t^{\mathsf{p}}\right)^{1/\mathsf{p}}$ by Theorem 14 in this case (note that both B_T and A_T are at most in the same order of T). Next, if $G_t = 0$, then these two bounds are hard to compare due to the extra term $D\left(\sum_{t=1}^T H_t^{\frac{2}{1-\nu}}\right)^{\frac{1-\nu}{2(1+\nu)}} [C_T(\boldsymbol{x})]^{\frac{\nu}{1+\nu}}$ in Theorem 15. But as one will see later, in convex optimization, Theorem 15 possibly leads to a better rate than Theorem 14.

Proof. By Lemma 10 with $\boldsymbol{h}_t = \boldsymbol{g}_{t-1}$ and $\gamma_t = +\infty$, we have

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \lesssim D\sqrt{\sum_{t=1}^T \left\|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}\right\|^2}. \tag{78}$$

Let $\boldsymbol{\epsilon}_0 \triangleq \mathbf{0}$, we can find

$$\begin{aligned}
&\sum_{t=1}^T \left\|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}\right\|^2 \\
=&\sum_{t=1}^T \left\|\nabla \ell_t(\boldsymbol{x}_t) - \nabla \ell_t(\boldsymbol{x}_t^\star) + \nabla \ell_t(\boldsymbol{x}_t^\star) - \nabla \ell_{t-1}(\boldsymbol{x}_{t-1}^\star) + \nabla \ell_{t-1}(\boldsymbol{x}_{t-1}^\star) - \nabla \ell_{t-1}(\boldsymbol{x}_{t-1}) + \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{t-1}\right\|^2 \\
\lesssim&\sum_{t=1}^T \left\|\nabla \ell_t(\boldsymbol{x}_t^\star) - \nabla \ell_{t-1}(\boldsymbol{x}_{t-1}^\star)\right\|^2 + \sum_{t=1}^T \left\|\nabla \ell_t(\boldsymbol{x}_t) - \nabla \ell_t(\boldsymbol{x}_t^\star)\right\|^2 + \sum_{t=1}^T \left\|\boldsymbol{\epsilon}_t\right\|^2 \\
=&B_T + \sum_{t=1}^T \left\|\nabla \ell_t(\boldsymbol{x}_t) - \nabla \ell_t(\boldsymbol{x}_t^\star)\right\|^2 + \sum_{t=1}^T \left\|\boldsymbol{\epsilon}_t\right\|^2.
\end{aligned} \tag{79}$$

By Condition 2, we know $\left\|\nabla \ell_t(\boldsymbol{x}_t) - \nabla \ell_t(\boldsymbol{x}_t^\star)\right\|^2 \lesssim G_t^2 + H_t^{\frac{2}{1+\nu}} (\ell_t(\boldsymbol{x}_t) - \ell_t^\star)^{\frac{2\nu}{1+\nu}}$, which implies

$$\sum_{t=1}^T \left\|\nabla \ell_t(\boldsymbol{x}_t) - \nabla \ell_t(\boldsymbol{x}_t^\star)\right\|^2 \lesssim \sum_{t=1}^T G_t^2 + H_t^{\frac{2}{1+\nu}} (\ell_t(\boldsymbol{x}_t) - \ell_t^\star)^{\frac{2\nu}{1+\nu}}. \tag{80}$$

Combine (79) and (80) and use $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}, \forall a, b \ge 0$ to have

$$\begin{aligned}
\sqrt{\sum_{t=1}^T \left\|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}\right\|^2} &\lesssim \sqrt{B_T + \sum_{t=1}^T G_t^2} + \sqrt{\sum_{t=1}^T H_t^{\frac{2}{1+\nu}} (\ell_t(\boldsymbol{x}_t) - \ell_t^\star)^{\frac{2\nu}{1+\nu}}} + \sqrt{\sum_{t=1}^T \left\|\boldsymbol{\epsilon}_t\right\|^2} \\
&\le \sqrt{B_T + \sum_{t=1}^T G_t^2} + \sqrt{\sum_{t=1}^T H_t^{\frac{2}{1+\nu}} (\ell_t(\boldsymbol{x}_t) - \ell_t^\star)^{\frac{2\nu}{1+\nu}}} + \left(\sum_{t=1}^T \left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},
\end{aligned} \tag{81}$$

where the last step is due to $\left\|\cdot\right\|_2 \le \left\|\cdot\right\|_{\mathsf{p}}$ for any $\mathsf{p} \in [1, 2]$.

Therefore, by (78) and (81), the following inequality holds

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \lesssim D\sqrt{\sum_{t=1}^T H_t^{\frac{2}{1+\nu}} (\ell_t(\boldsymbol{x}_t) - \ell_t^\star)^{\frac{2\nu}{1+\nu}}} + D\sqrt{B_T + \sum_{t=1}^T G_t^2} + D\left(\sum_{t=1}^T \left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}.$$

We use Hölder's inequality to bound

$$\sum_{t=1}^T H_t^{\frac{2}{1+\nu}} (\ell_t(\boldsymbol{x}_t) - \ell_t^\star)^{\frac{2\nu}{1+\nu}} \le \left(\sum_{t=1}^T H_t^{\frac{2}{1-\nu}}\right)^{\frac{1-\nu}{1+\nu}} \left(\sum_{t=1}^T \ell_t(\boldsymbol{x}_t) - \ell_t^\star\right)^{\frac{2\nu}{1+\nu}},$$

which implies

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \lesssim D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left(\sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \ell_t^\star \right)^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}}.$$

Now take expectations on both sides of the above inequality and use the fact $\mathbb{E}\left[\left(\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right] \leq \left(\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}} \leq D\left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}$ (due to Hölder's inequality) to find for $\mathsf{A} = \mathsf{OAdaGrad}$

$$\begin{aligned}
\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] &\lesssim D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \mathbb{E}\left[\left(\sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \ell_t^\star \right)^{\frac{\nu}{1+\nu}} \right] + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}} \\
&\leq D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left(\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \ell_t^\star \right] \right)^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}} \\
&= D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left(\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] + C_T(\boldsymbol{x}) \right)^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}}, \qquad (82)
\end{aligned}$$

where the second step is due to Hölder's inequality.

If $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \leq 0$, we are done. Hence, we only need to consider the case $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \geq 0$. If $\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \leq C_T(\boldsymbol{x})$, then (82) implies

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left[C_T(\boldsymbol{x})\right]^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}}.$$

Otherwise, we have

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left(\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \right)^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}},$$

which implies

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim D^{1+\nu} \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2}} + D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left[C_T(\boldsymbol{x})\right]^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}}.$$

So in all three cases, we have

$$\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x})\right] \lesssim D^{1+\nu} \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2}} + D \left(\sum_{t=1}^{T} H_t^{\frac{2}{1-\nu}} \right)^{\frac{1-\nu}{2(1+\nu)}} \left[C_T(\boldsymbol{x})\right]^{\frac{\nu}{1+\nu}} + D \sqrt{B_T + \sum_{t=1}^{T} G_t^2 + D \left(\sum_{t=1}^{T} \sigma_t^{\mathsf{p}} \right)^{\frac{1}{\mathsf{p}}}}.$$

□

G More Applications

Based on the results for OAdaGrad given in Appendix F, we provide more applications in this section.

G.1 General Nonsmooth Convex Optimization

With Theorems 14 and 15, we immediately obtain the following convergence rates for convex optimization. The proof of which is omitted to save space.

Corollary 8. *Under Assumption 3 for $\ell_t(\boldsymbol{x}) = F(\boldsymbol{x})$ (meaning that $G_t = G$, $H_t = H$, and $\sigma_t = \sigma$) and let $\boldsymbol{x}^\star \triangleq \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ and $\bar{\boldsymbol{x}}_T \triangleq \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t$:*

- *considering the same setting for* $\mathsf{A} = \mathsf{OAdaGrad}$ *as in Theorem 14, we have*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}^\star)\right] \le \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}^\star)\right]}{T} \lesssim \frac{\left\|\nabla F(\boldsymbol{x}^\star)\right\| D}{T} + \frac{HD^{1+\nu}}{T^{\frac{1}{2-\nu}}} + \frac{GD}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}.$$

- *replacing the third point in Assumption 3 by Condition 2 and considering the same setting for* $\mathsf{A} = \mathsf{OAdaGrad}$ *as in Theorem 15, we have*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}_T) - F(\boldsymbol{x}^\star)\right] \le \frac{\mathbb{E}\left[\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{x}^\star)\right]}{T} \lesssim \frac{\left\|\nabla F(\boldsymbol{x}^\star)\right\| D}{T} + \frac{HD^{1+\nu}}{T^{\frac{1+\nu}{2}}} + \frac{GD}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}.$$

If we consider Assumption 1, (i.e., $H = 0$ and $\left\|\nabla F(\boldsymbol{x})\right\| \le G$), Corollary 8 is as fast as Corollary 1 and hence optimal. Under the same assumption used in Corollary 7 (i.e., $G = 0$ and $\nu = 1$), both rates degenerates to $\frac{\left\|\nabla F(\boldsymbol{x}^\star)\right\| D}{T} + \frac{HD^2}{T} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}$ and are faster than the bound $\frac{HD^2}{T} + \frac{\left\|\nabla F(\boldsymbol{x}^\star)\right\| D}{\sqrt{T}} + \frac{\sigma D}{T^{1-\frac{1}{\mathsf{p}}}}$ given in Corollary 7. If we specialize in globally Hölder smooth functions (i.e., $\left\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\right\| \le H\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^\nu, \forall, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$), then both bounds are new under heavy tails and the second one is faster.

G.2 General Nonsmooth Nonconvex Optimization

In this section, we move back to the nonsmooth nonconvex optimization studied before in Section 4. But instead of assuming F is Lipschitz as in Assumption 2, we will use the following general Assumption 4. Note that a function satisfying Assumption 2 with a Lipschitz parameter G also fits Assumption 4 for $H = 0$.

Assumption 4. *We consider the following series of assumptions:*

- *The objective F is lower bounded by $F_\star \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) \in \mathbb{R}$.*
- *F is differentiable and well-behaved, i.e., $F(\boldsymbol{x}) - F(\boldsymbol{y}) = \int_0^1 \langle \nabla F(\boldsymbol{y} + t(\boldsymbol{x}-\boldsymbol{y})), \boldsymbol{x}-\boldsymbol{y}\rangle\,\mathrm{d}t$.*
- *F is (G, H, ν)-general nonsmooth on $\mathbb{R}^d$, i.e., there exists $G \ge 0$, $H \ge 0$ and $\nu \in (0,1]$ such that $G + H > 0$ and $\left\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\right\| \le 2G + H\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^\nu, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.*
- *Given $\boldsymbol{z}_t \in \mathbb{R}^d$ at the t-th iteration, one can query $\boldsymbol{g}_t \in \mathbb{R}^d$ satisfying $\mathbb{E}\left[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right] = \nabla F(\boldsymbol{z}_t)$ and $\mathbb{E}\left[\left\|\boldsymbol{\epsilon}_t\right\|^{\mathsf{p}}\right] \le \sigma^{\mathsf{p}}$ for some $\mathsf{p} \in (1,2]$ and $\sigma \ge 0$, where $\mathcal{F}_t$ denotes the natural filtration and $\boldsymbol{\epsilon}_t \triangleq \boldsymbol{g}_t - \nabla F(\boldsymbol{z}_t)$ is the stochastic noise.*

Our goal is still to find a (δ, ϵ)-stationary point, but under Assumption 4 this time. Fortunately, we can still start from the $\mathsf{O2NC}$ framework (Algorithm 4), since the proof of Theorem 4 only relies on the first two and the last one conditions in Assumption 2, which are the same as Assumption 4. In other words, Theorem 4 holds under this more general Assumption 4.

G.2.1 OAdaGrad with Reset

Algorithm 6 OAdaGrad with Reset (OAdaGradR)

Input: initial point $\boldsymbol{x}_1 \in \mathcal{B}^d(D)$, parameter $D > 0$.
Set $\boldsymbol{g}_0 = \mathbf{0}$ and $V_0 = 0$
for $n = 1$ **to** KT **do**
  $V_n = V_{n-1}\mathbb{1}[n \mod T \neq 1] + \left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2$
  $\eta_n = \sqrt{2}D/\sqrt{V_n}$
  $\boldsymbol{x}_{n+1} = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{B}^d(D)} \langle 2\boldsymbol{g}_n - \boldsymbol{g}_{n-1}, \boldsymbol{x}\rangle + \frac{\|\boldsymbol{x}-\boldsymbol{x}_n\|^2}{2\eta_n}$
  if $n+1 \mod T = 1$ **do**
    $\boldsymbol{x}_{n+1} = -D\frac{\boldsymbol{g}_n}{\|\boldsymbol{g}_n\|}$
  end if
end for

With Theorem 4 on hand, our next task is naturally to find an online learning algorithm A that has a proper K-shifting regret under Assumption 4 (especially the third point in which). Thanks to the framework presented in Appendix F, we already have some clues, i.e., employing an optimistic algorithm. In fact, such an idea has also been studied in [7] but only for the deterministic smooth case, i.e., $\sigma = G = 0$ and $\nu = 1$ in Assumption 4.

To handle the more general case, particularly including heavy-tailed noise, we first present a new method called OAdaGrad with Reset (OAdaGradR) in Algorithm 6, which can be viewed as running OAdaGrad for T iterations with the hint $\boldsymbol{h}_t = \boldsymbol{g}_{t-1}$ and then resetting in total of K times.

We clarify that the idea of reset is not new, which was originally suggested by [7], as also used previously in Theorem 5. However, the step of how to reset in OAdaGradR, i.e., $\boldsymbol{x}_{n+1} = -D\frac{\boldsymbol{g}_n}{\|\boldsymbol{g}_n\|}$ if $n+1 \mod T = 1$, is critical and novel as far as we know. Why is it important? This is because the hint $\boldsymbol{h}_t$ is not reset to $\mathbf{0}$. Hence, if we reset $\boldsymbol{x}_{n+1}$ to an arbitrary point in $\mathcal{B}^d(D)$, then one can imagine that we will face a redundant term $\sum_{k=1}^K \langle \boldsymbol{h}_{(k-1)T}, \boldsymbol{x}_{(k-1)T+1} - \boldsymbol{v}_k\rangle$ in total (see Lemma 10), which is however undesired. Instead, our specially designed way of resetting $\boldsymbol{x}_{n+1}$ can resolve this potential issue, as reflected in the following Lemma 11.

Lemma 11. *For any initial point $\boldsymbol{x}_1 \in \mathcal{B}^d(D)$ and any sequence $\boldsymbol{v}_1, \dots, \boldsymbol{v}_K$ satisfying $\boldsymbol{v}_k \in \mathcal{B}^d(D), \forall k \in [K]$, the online learning algorithm* $\mathsf{A} = \mathsf{OAdaGradR}$ *(Algorithm 6) guarantees*

$$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_K) \lesssim D\sum_{k=1}^K \sqrt{\sum_{n=(k-1)T+1}^{kT} \left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2},$$

where $\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_K)$ is the K-shifting regret defined in (9).

Proof. Given $k \in [K]$, for $n \in \{(k-1)T+1, \dots, kT\}$, OAdaGradR (Algorithm 6) is the same as running OAdaGrad (Algorithm 5) for T iterations with the feasible set $\mathcal{X} = \mathcal{B}^d(D)$ (which implies $\sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{y}\| \leq 2D$), initial point $\boldsymbol{x}_{(k-1)T+1}$, hint sequence $\boldsymbol{h}_t = \boldsymbol{g}_{(k-1)T+t-1}$, stepsize $\eta = 2D/\sqrt{2}$ and $\gamma_t = +\infty$, and the stochastic gradient sequence $\boldsymbol{g}_{(k-1)T+t}$. Therefore, Lemma 10 implies that for any $\boldsymbol{v}_k \in \mathcal{B}^d(D)$,

$$\begin{aligned}\sum_{n=(k-1)T+1}^{kT} \langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k\rangle &\lesssim \left\langle \boldsymbol{g}_{(k-1)T}, \boldsymbol{x}_{(k-1)T+1} - \boldsymbol{v}_k\right\rangle + D\sqrt{\sum_{t=1}^T \left\|\boldsymbol{g}_{(k-1)T+t} - \boldsymbol{g}_{(k-1)T+t-1}\right\|^2}. \\ &= \left\langle \boldsymbol{g}_{(k-1)T}, \boldsymbol{x}_{(k-1)T+1} - \boldsymbol{v}_k\right\rangle + D\sqrt{\sum_{n=(k-1)T+1}^{kT} \left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2} \end{aligned} \tag{83}$$

Observe that if $k = 1$, we have

$$\left\langle \boldsymbol{g}_{(k-1)T}, \boldsymbol{x}_{(k-1)T+1} - \boldsymbol{v}_k\right\rangle = \langle \boldsymbol{g}_0, \boldsymbol{x}_1 - \boldsymbol{v}_1\rangle = \langle \mathbf{0}, \boldsymbol{x}_1 - \boldsymbol{v}_1\rangle = 0.$$

If $k \neq 1$, we use $\boldsymbol{x}_{(k-1)T+1} = -D\frac{\boldsymbol{x}_{(k-1)T}}{\|\boldsymbol{x}_{(k-1)T}\|}$ and $\|\boldsymbol{v}_k\| \leq D$ to have

$$\left\langle \boldsymbol{g}_{(k-1)T}, \boldsymbol{x}_{(k-1)T+1} - \boldsymbol{v}_k \right\rangle = -D\left\|\boldsymbol{g}_{(k-1)T}\right\| + \left\langle \boldsymbol{g}_{(k-1)T}, -\boldsymbol{v}_k \right\rangle \leq 0.$$

Thus there is always

$$\left\langle \boldsymbol{g}_{(k-1)T}, \boldsymbol{x}_{(k-1)T+1} - \boldsymbol{v}_k \right\rangle \leq 0. \tag{84}$$

Finally, we combine (83) and (84) to know for any $k \in [K]$,

$$\sum_{n=(k-1)T+1}^{kT} \langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k \rangle \lesssim D\sqrt{\sum_{n=(k-1)T+1}^{kT} \left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2}, \forall \boldsymbol{v}_k \in \mathcal{B}^d(D),$$

sum up which from $k = 1$ to K to obtain the following desired result

$$\mathsf{R}_T^{\mathsf{A}}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) = \sum_{k=1}^{K} \sum_{n=(k-1)T+1}^{kT} \langle \boldsymbol{g}_n, \boldsymbol{x}_n - \boldsymbol{v}_k \rangle \lesssim D\sum_{k=1}^{K}\sqrt{\sum_{n=(k-1)T+1}^{kT} \left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2}.$$

□

G.2.2 Convergence Rates

Equipped with Lemma 11 for OAdaGradR, we are ready to show convergence rates under the new Assumption 4. First, we prove the following Theorem 16, which can be viewed as a generalization of Theorem 5.

Theorem 16. *Under Assumption 4 and let* $\nabla \triangleq \|\nabla F(\boldsymbol{y}_0)\|$, $\Delta \triangleq F(\boldsymbol{y}_0) - F_\star$, *and* $\bar{\boldsymbol{z}}_k \triangleq \frac{1}{T}\sum_{n=(k-1)T+1}^{kT} \boldsymbol{z}_n, \forall k \in [K]$*, setting* $\mathsf{A} = \mathsf{OAdaGradR}$ *(Algorithm 6) in* $\mathsf{O2NC}$ *(Algorithm 4) with* $D = \delta/T$*, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta\right] \lesssim \frac{\nabla}{KT} + \frac{\Delta}{\delta K} + \frac{H\delta^\nu}{T^{\frac{1}{2}+\nu}} + \frac{G}{\sqrt{T}} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.$$

Proof. By Theorem 4 and Lemma 11, there is

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{1}{K}\left\|\frac{1}{T}\sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n)\right\|\right] \lesssim \frac{F(\boldsymbol{y}_0) - F_\star}{DKT} + \frac{\sum_{k=1}^{K}\mathbb{E}\left[\sqrt{\sum_{n=(k-1)T+1}^{kT}\left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2}\right]}{KT} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}. \tag{85}$$

We first lower bound the L.H.S. of (85). Same as (45), we still have

$$\boldsymbol{z}_n \in \mathcal{B}^d(\bar{\boldsymbol{z}}_k, \delta), \forall n \in \{(k-1)T+1, \ldots, kT\}.$$

By the definition of $\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta$ (see Definition 1), there is

$$\|\nabla F(\bar{\boldsymbol{z}}_k)\|_\delta \leq \left\|\frac{1}{T}\sum_{n=(k-1)T+1}^{kT} \nabla F(\boldsymbol{z}_n)\right\|. \tag{86}$$

Next, we upper bound the R.H.S. of (85). Before moving on, we list two facts that will be frequently used later, i.e., for any $n \in [KT]$,

$$\|\boldsymbol{x}_n\| \leq D \quad \text{and} \quad s_n \in [0,1]. \tag{87}$$

Now, let us upper bound $\left\|\boldsymbol{g}_n - \boldsymbol{g}_{n-1}\right\|^2$ on the R.H.S. of (85) as follows.

- $n=1$. In this case, we recall $\boldsymbol{g}_0=\mathbf{0}$ and hence have

$$
\begin{aligned}
\left\|\boldsymbol{g}_n-\boldsymbol{g}_{n-1}\right\|^2 &= \left\|\boldsymbol{g}_1\right\|^2=\left\|\boldsymbol{\epsilon}_1+\nabla F(\boldsymbol{z}_1)-\nabla F(\boldsymbol{y}_0)+\nabla F(\boldsymbol{y}_0)\right\|^2\\
&\lesssim \left\|\boldsymbol{\epsilon}_1\right\|^2+\left\|\nabla F(\boldsymbol{z}_1)-\nabla F(\boldsymbol{y}_0)\right\|^2+\left\|\nabla F(\boldsymbol{y}_0)\right\|^2\\
&\overset{\text{Assumption 4}}{\lesssim} \left\|\boldsymbol{\epsilon}_1\right\|^2+G^2+H^2\left\|\boldsymbol{z}_1-\boldsymbol{y}_0\right\|^{2\nu}+\left\|\nabla F(\boldsymbol{y}_0)\right\|^2\\
&\overset{(a)}{\lesssim} \left\|\boldsymbol{\epsilon}_1\right\|^2+G^2+H^2D^{2\nu}+\left\|\nabla F(\boldsymbol{y}_0)\right\|^2, \qquad (88)
\end{aligned}
$$

  where (a) is by $\left\|\boldsymbol{z}_1-\boldsymbol{y}_0\right\|=s_1\left\|\boldsymbol{x}_1\right\|\overset{(87)}{\le} D$.

- $n\neq 1$. In this case, we have

$$
\begin{aligned}
\left\|\boldsymbol{g}_n-\boldsymbol{g}_{n-1}\right\|^2 &= \left\|\boldsymbol{\epsilon}_n-\boldsymbol{\epsilon}_{n-1}+\nabla F(\boldsymbol{z}_n)-\nabla F(\boldsymbol{z}_{n-1})\right\|^2\\
&\lesssim \left\|\boldsymbol{\epsilon}_n\right\|^2+\left\|\boldsymbol{\epsilon}_{n-1}\right\|^2+\left\|\nabla F(\boldsymbol{z}_n)-\nabla F(\boldsymbol{z}_{n-1})\right\|^2\\
&\overset{\text{Assumption 4}}{\lesssim} \left\|\boldsymbol{\epsilon}_n\right\|^2+\left\|\boldsymbol{\epsilon}_{n-1}\right\|^2+G^2+H^2\left\|\boldsymbol{z}_n-\boldsymbol{z}_{n-1}\right\|^{2\nu}\\
&\overset{(b)}{\lesssim} \left\|\boldsymbol{\epsilon}_n\right\|^2+\left\|\boldsymbol{\epsilon}_{n-1}\right\|^2+G^2+H^2D^{2\nu}, \qquad (89)
\end{aligned}
$$

  where (b) is by $\left\|\boldsymbol{z}_n-\boldsymbol{z}_{n-1}\right\|=\left\|s_n\boldsymbol{x}_n+(1-s_{n-1})\boldsymbol{x}_{n-1}\right\|\le s_n\left\|\boldsymbol{x}_n\right\|+(1-s_{n-1})\left\|\boldsymbol{x}_{n-1}\right\|\overset{(87)}{\le} 2D$.

Thus, we can find for any $k\in[K]$,

$$
\sum_{n=(k-1)T+1}^{kT}\left\|\boldsymbol{g}_n-\boldsymbol{g}_{n-1}\right\|^2 \overset{(88),(89)}{\lesssim} \left\|\nabla F(\boldsymbol{y}_0)\right\|^2\mathbb{1}\left[k=1\right]+\left(G^2+H^2D^{2\nu}\right)T+\sum_{n=(k-1)T}^{kT}\left\|\boldsymbol{\epsilon}_n\right\|^2,
$$

where $\boldsymbol{\epsilon}_0\triangleq\mathbf{0}$ for simplicity. As such, we obtain

$$
\begin{aligned}
\sqrt{\sum_{n=(k-1)T+1}^{kT}\left\|\boldsymbol{g}_n-\boldsymbol{g}_{n-1}\right\|^2} &\lesssim \left\|\nabla F(\boldsymbol{y}_0)\right\|\mathbb{1}\left[k=1\right]+\left(HD^{\nu}+G\right)\sqrt{T}+\sqrt{\sum_{n=(k-1)T}^{kT}\left\|\boldsymbol{\epsilon}_n\right\|^2}\\
&\le \left\|\nabla F(\boldsymbol{y}_0)\right\|\mathbb{1}\left[k=1\right]+\left(HD^{\nu}+G\right)\sqrt{T}+\left(\sum_{n=(k-1)T}^{kT}\left\|\boldsymbol{\epsilon}_n\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}},
\end{aligned}
$$

where the last step is by $\left\|\cdot\right\|_2\le\left\|\cdot\right\|_{\mathsf{p}}$ for $\mathsf{p}\in[1,2]$. By Hölder's inequality, there is

$$
\mathbb{E}\left[\left(\sum_{n=(k-1)T}^{kT}\left\|\boldsymbol{\epsilon}_n\right\|^{\mathsf{p}}\right)^{\frac{1}{\mathsf{p}}}\right]\le\left(\sum_{n=(k-1)T}^{kT}\mathbb{E}\left[\left\|\boldsymbol{\epsilon}_n\right\|^{\mathsf{p}}\right]\right)^{\frac{1}{\mathsf{p}}}\overset{\text{Assumption 4}}{\lesssim}\sigma(T+1)^{\frac{1}{\mathsf{p}}}.
$$

Thus, we know

$$
\mathbb{E}\left[\sqrt{\sum_{n=(k-1)T+1}^{kT}\left\|\boldsymbol{g}_n-\boldsymbol{g}_{n-1}\right\|^2}\right]\lesssim\left\|\nabla F(\boldsymbol{y}_0)\right\|\mathbb{1}\left[k=1\right]+\left(HD^{\nu}+G\right)\sqrt{T}+\sigma T^{\frac{1}{\mathsf{p}}},\forall k\in\left[K\right]. \qquad (90)
$$

Finally, we plug (86) and (90) back into (85), then use $D=\delta/T$, $\nabla=\left\|\nabla F(\boldsymbol{y}_0)\right\|$, and $\Delta=F(\boldsymbol{y}_0)-F_\star$ to have

$$
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\bar{\boldsymbol{z}}_k)\right\|_\delta\right]\lesssim\frac{\left\|\nabla F(\boldsymbol{y}_0)\right\|}{KT}+\frac{F(\boldsymbol{y}_0)-F_\star}{\delta K}+\frac{H\delta^{\nu}}{T^{\frac{1}{2}+\nu}}+\frac{G}{\sqrt{T}}+\frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}}.
$$

□

Similar to before, we can consider two situations where problem-dependent parameters are known or unknown, corresponding to the following Corollaries 9 and 10, respectively. To save space, the proofs are omitted since they can be easily checked.

Corollary 9. *Under the same setting in Theorem 16, suppose we have $N \geq 2$ stochastic gradient budgets, taking $K = \lfloor N/T \rfloor$ and $T = \lceil N/2 \rceil \wedge \left(\left\lceil \left(\frac{\delta^{1+\nu} HN}{\Delta} \right)^{\frac{2}{3+2\nu}} \right\rceil \vee \left\lceil \left(\frac{\delta GN}{\Delta} \right)^{\frac{2}{3}} \right\rceil \vee \left\lceil \left(\frac{\delta \sigma N}{\Delta} \right)^{\frac{\mathsf{p}}{2\mathsf{p}-1}} \right\rceil \right)$, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{y}}_k)\|_\delta\right] \lesssim \frac{\nabla}{N} + \frac{H\delta^\nu}{N^{\frac{1}{2}+\nu}} + \frac{G}{\sqrt{N}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \frac{\Delta}{\delta N} + \frac{H^{\frac{2}{3+2\nu}}\Delta^{\frac{1+2\nu}{3+2\nu}}}{\delta^{\frac{1}{3+2\nu}}N^{\frac{1+2\nu}{3+2\nu}}} + \frac{G^{\frac{2}{3}}\Delta^{\frac{1}{3}}}{(\delta N)^{\frac{1}{3}}} + \frac{\sigma^{\frac{\mathsf{p}}{2\mathsf{p}-1}}\Delta^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}{(\delta N)^{\frac{\mathsf{p}-1}{2\mathsf{p}-1}}}.$$

Corollary 10. *Under the same setting in Theorem 16, suppose we have $N \geq 2$ stochastic gradient budgets, taking $K = \lfloor N/T \rfloor$ and $T = \lceil N/2 \rceil \wedge \left\lceil (\delta N)^{\frac{2}{3}} \right\rceil$, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{y}}_k)\|_\delta\right] \lesssim \frac{\nabla}{N} + \frac{H\delta^\nu}{N^{\frac{1}{2}+\nu}} + \frac{H}{\delta^{\frac{2-\nu}{5}}N^{\frac{1+2\nu}{5}}} + \frac{\Delta}{(\delta N) \wedge (\delta N)^{\frac{1}{3}}} + \frac{G}{\sqrt{N} \wedge (\delta N)^{\frac{1}{3}}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}} \wedge (\delta N)^{\frac{2(\mathsf{p}-1)}{3\mathsf{p}}}}.$$

G.2.3 A Special Case: Hölder Smooth Nonconvex Functions

We now consider a special case of $G = 0$ in Assumption 4, meaning that F is Hölder smooth. Now, due to the smoothness, finding an ϵ-stationary point instead of a (δ, ϵ)-stationary point is more reasonable. In the following Lemma 12, we connect these two notions. Especially, when $\nu = 1$, Lemma 12 recovers Proposition 14 of [7].

Lemma 12. *If F is (H,ν)-Hölder smooth, i.e., there exists $H > 0$ and $\nu \in (0,1]$ such that $\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq H\|\boldsymbol{x} - \boldsymbol{y}\|^\nu, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, then $\|\nabla F(\boldsymbol{x})\| \leq \|\nabla F(\boldsymbol{x})\|_\delta + H\delta^\nu, \forall \boldsymbol{x} \in \mathbb{R}^d, \delta > 0$.*

Proof. For any fixed $\boldsymbol{x} \in \mathbb{R}^d$ and $\delta, \epsilon > 0$, by the definition of $\|\nabla F(\boldsymbol{x})\|_\delta$, there exists a finite set $S \subset \mathcal{B}^d(\boldsymbol{x}, \delta)$ such that $\frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\boldsymbol{y} = \boldsymbol{x}$ and $\left\|\frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\nabla F(\boldsymbol{y})\right\| \leq \|\nabla F(\boldsymbol{x})\|_\delta + \epsilon$. Hence, we know

$$\begin{aligned}
\|\nabla F(\boldsymbol{x})\| &\leq \left\|\frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\nabla F(\boldsymbol{y})\right\| + \left\|\nabla F(\boldsymbol{x}) - \frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\nabla F(\boldsymbol{y})\right\| \\
&\leq \|\nabla F(\boldsymbol{x})\|_\delta + \epsilon + \left\|\nabla F(\boldsymbol{x}) - \frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\nabla F(\boldsymbol{y})\right\| \\
&\leq \|\nabla F(\boldsymbol{x})\|_\delta + \epsilon + \frac{1}{|S|}\sum_{\boldsymbol{y}\in S}\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \\
&\leq \|\nabla F(\boldsymbol{x})\|_\delta + \epsilon + \frac{H}{|S|}\sum_{\boldsymbol{y}\in S}\|\boldsymbol{x} - \boldsymbol{y}\|^\nu \leq \|\nabla F(\boldsymbol{x})\|_\delta + \epsilon + H\delta^\nu.
\end{aligned}$$

Take $\epsilon \to 0$ to conclude. □

Armed with Lemma 12, we can prove the following Theorem 17.

Theorem 17. *Under Assumption 4 (with $G = 0$) and let $\nabla \triangleq \|\nabla F(\boldsymbol{y}_0)\|$, $\Delta \triangleq F(\boldsymbol{y}_0) - F_\star$, and $\bar{\boldsymbol{z}}_k \triangleq \frac{1}{T}\sum_{n=(k-1)T+1}^{kT}\boldsymbol{z}_n, \forall k \in [K]$, setting $\mathsf{A} = \mathsf{OAdaGradR}$ (Algorithm 6) in $\mathsf{O2NC}$ (Algorithm 4) with $D = \delta/T$, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{z}}_k)\|\right] \lesssim \frac{\nabla}{KT} + \frac{\Delta}{\delta K} + \frac{\sigma}{T^{1-\frac{1}{\mathsf{p}}}} + H\delta^\nu.$$

Proof. We first invoke Lemma 12 and Theorem 16 with $G = 0$, then use $T \geq 1$ to conclude. $\square$

Note that now δ should also be viewed as a parameter decided by the user. Therefore, we will also choose the value of δ in the following two corollaries, corresponding to the cases where problem-dependent parameters are known and unknown, respectively. Again, the proofs are omitted to save space, and they are easy to check.

Corollary 11. *Under the same setting in Theorem 17, let* $r \triangleq \mathsf{p}\nu + (\mathsf{p}-1)(1+\nu)$*, suppose we have* $N \geq 2$ *stochastic gradient budgets, taking* $K = \lfloor N/T \rfloor$*,* $T = \lceil N/2 \rceil \wedge \left\lceil \frac{\sigma^{\frac{\mathsf{p}(1+\nu)}{r}} N^{\frac{\mathsf{p}\nu}{r}}}{H^{\frac{\mathsf{p}}{r}} \Delta^{\frac{\mathsf{p}\nu}{r}}} + 1 \right\rceil$*, and* $\delta = \left(\frac{\Delta T}{HN}\right)^{\frac{1}{1+\nu}}$*, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{y}}_k)\|\right] \lesssim \frac{\nabla}{N} + \frac{H^{\frac{1}{1+\nu}}\Delta^{\frac{\nu}{1+\nu}}}{N^{\frac{\nu}{1+\nu}}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \frac{\sigma^{\frac{\mathsf{p}\nu}{r}} H^{\frac{\mathsf{p}-1}{r}} \Delta^{\frac{(\mathsf{p}-1)\nu}{r}}}{N^{\frac{(\mathsf{p}-1)\nu}{r}}}.$$

Corollary 12. *Under the same setting in Theorem 17, suppose we have* $N \geq 2$ *stochastic gradient budgets, taking* $K = \lfloor N/T \rfloor$*,* $T = \left\lceil \sqrt{N} \right\rceil$*, and* $\delta = 1/N^{1/4}$*, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\boldsymbol{y}}_k)\|\right] \lesssim \frac{\nabla}{N} + \frac{\Delta}{N^{\frac{1}{4}}} + \frac{H}{N^{\frac{\nu}{4}}} + \frac{\sigma}{N^{\frac{\mathsf{p}-1}{2\mathsf{p}}}}.$$

As a sanity check, when $\nu = 1$ (i.e., the standard smooth case), using the fact $\nabla = \|\nabla F(\boldsymbol{y}_0)\| \lesssim \sqrt{H(F(\boldsymbol{y}_0) - F_\star)} = \sqrt{H\Delta}$, Corollary 11 reduces to a rate $\sqrt{\frac{H\Delta}{N}} + \frac{\sigma}{N^{1-\frac{1}{\mathsf{p}}}} + \left(\frac{\sigma^{\frac{\mathsf{p}}{\mathsf{p}-1}} H\Delta}{N}\right)^{\frac{\mathsf{p}-1}{3\mathsf{p}-2}}$ and Corollary 12 gives a rate $\frac{\Delta+H}{N^{\frac{1}{4}}} + \frac{\sigma}{N^{\frac{\mathsf{p}-1}{2\mathsf{p}}}}$, both of which match the best possible results in the respective situations [14, 23].

For general $\nu \in (0, 1)$, as far as we know, no previous works consider heavy-tailed noise. Hence, both these Corollaries are the first and new.

H Algebraic Facts

We give three useful algebraic facts in this section.

Fact 1. *For any* $T \in \mathbb{N}$ *and* $a \in (0, 1)$*, there is*

$$\sum_{t=1}^{T-1} \frac{\sum_{s=t+1}^{T} s^a}{t(T-t)^2} \lesssim \frac{1 + \log T}{T^{1-a}}.$$

Proof. Note that $\sum_{s=t+1}^{T} s^a \leq (T-t)T^a$, which implies

$$\sum_{t=1}^{T-1} \frac{\sum_{s=t+1}^{T} s^a}{t(T-t)^2} \leq \sum_{t=1}^{T-1} \frac{T^a}{t(T-t)} = \frac{1}{T^{1-a}} \sum_{t=1}^{T-1} \frac{1}{t} + \frac{1}{T-t} = \frac{2}{T^{1-a}} \sum_{t=1}^{T-1} \frac{1}{t} \lesssim \frac{1 + \log T}{T^{1-a}}.$$

$\square$

Fact 2. *Given* $2 \leq N \in \mathbb{N}$*,* $K = \lfloor N/T \rfloor$ *and* $T \in \mathbb{N}$ *satisfying* $T \leq \lceil N/2 \rceil$*, there is* $KT \geq N/4$*.*

Proof. Note that $KT = \lfloor N/T \rfloor T \geq N - T \geq (N-1)/2 \geq N/4$. $\square$

Fact 3. *Given* $\mathsf{p} \in (1, 2]$ *and* $q \in (0, 1)$*, there are* $q^{\mathsf{p}-1} + (1-q)^{\mathsf{p}-1} \leq 2^{2-\mathsf{p}}$ *and* $1 - q \leq \frac{1-q^{\mathsf{p}-1}}{\mathsf{p}-1}$*.*

Proof. Note that $x^{\mathsf{p}-1}$ is concave when $\mathsf{p} \in (1, 2]$, we hence have $\frac{q^{\mathsf{p}-1}+(1-q)^{\mathsf{p}-1}}{2} \leq \left(\frac{q+1-q}{2}\right)^{\mathsf{p}-1} \Rightarrow q^{\mathsf{p}-1} + (1-q)^{\mathsf{p}-1} \leq 2^{2-\mathsf{p}}$. Next, let $x = 1 - q \in (0, 1)$, we have

$$1 - q \leq \frac{1 - q^{\mathsf{p}-1}}{\mathsf{p}-1} \Leftrightarrow (1-x)^{\mathsf{p}-1} \leq 1 - (\mathsf{p}-1)x,$$

which is true by Bernoulli's inequality. $\square$